ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ στα

ΠΟΛΥΠΛΟΚΑ ΣΥΣΤΗΜΑΤΑ και ΔΙΚΤΥΑ

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ
ΤΜΗΜΑ ΓΕΩΛΟΓΙΑΣ
ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Τίτλος Εργασίας

Ανάλυση στατιστικής σημαντικότητας του modularity σε δίκτυα χρονοσειρών

Statistical significance of modularity measures on time series networks

Χριστόπουλος Χ. Γεώργιος

ΕΠΙΒΛΕΠΩΝ: Κουγιουμτζής Δημήτρης, Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Η/Υ, Α.Π.Θ

Θεσσαλονίκη , Δεκέμβριος 2019

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ στα
ΠΟΛΥΠΛΟΚΑ ΣΥΣΤΗΜΑΤΑ και ΔΙΚΤΥΑ
ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ
ΤΜΗΜΑ ΓΕΩΛΟΓΙΑΣ
ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Τίτλος Εργασίας

Ανάλυση στατιστικής σημαντικότητας του modularity σε δίκτυα χρονοσειρών

Statistical significance of modularity measures on time series networks

Χριστόπουλος Χ. Γεώργιος

ΕΠΙΒΛΕΠΩΝ: Κουγιουμτζής Δημήτρης, Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Η/Υ, Α.Π.Θ

Εγκρίθηκε από την Τριμελή Εξεταστική Επιτροπή την 19η Δεκεμβρίου 2019.

Κουγιουμτζής Δημήτρης, Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Η/Υ, Α.Π.Θ	Αντωνίου Ιωάννης, Καθηγητής, Τμήμα Μαθηματικών, Α.Π.Θ.	Καραγιάννης Βασίλειος, Μέλος ΕΔΙΠ, Τμήμα Μαθηματικών, ΑΠΘ.
……	……	……

Θεσσαλονίκη , Δεκέμβριος 2019

……..

Χριστόπουλος Χ. Γεώργιος
Πτυχιούχος Μαθηματικός Α.Π.Θ.

ΠΕΡΙΛΗΨΗ

Η χρήση δικτύων αποτελεί μια από τις δημοφιλέστερες μεθόδους στην ανάλυση χρονοσειρών. Στην παρούσα εργασία αξιολογείται η ισχύς και το μέγεθος του ελέγχου υπόθεσης, κατά τον οποίο η μηδενική υπόθεση H_0 ορίζει το δίκτυο συσχέτισης ως τυχαίο και χωρίς την παρουσία δομής κοινοτήτων. Το στατιστικό που χρησιμοποιήθηκε για την αξιολόγηση του ελέγχου υπόθεσης είναι η συνάρτηση modularity. Χρησιμοποιήθηκαν έξι διαφορετικές μέθοδοι τυχαιοποίησης δικτύου και σε πολλά διαφορετικά σενάρια πραγματοποίησης πολυμεταβλητών χρονοσειρών. Τέλος κατά την σύγκριση των μεθόδων τυχαιοποίησης δικτύων αναδείχθηκε η ανωτερότητα των μεθόδων που πραγματοποιούν την τυχαιοποίηση στις χρονοσειρές σε σύγκριση με αυτές που πραγματοποιούν τυχαιοποίηση απευθείας στις συνδέσεις του δικτύου συσχέτισης.

Αξίζει να σημειωθεί ότι ο έλεγχος υπόθεσης για την έλλειψη δομής κοινοτήτων εφαρμόζεται και σε πραγματικά δεδομένα. Εφαρμόστηκε σε δίκτυα συσχέτισης που δημιουργούνται από τα οικονομικά δεδομένα της εταιρίας Morgan Stanley Capital International και περιγράφουν την κεφαλαιοποίηση στην διεθνή αγορά κεφαλαίου 55 χωρών κατά το χρονικό διάστημα Μάρτιος 2004 έως Μάρτιος 2009.

ΛΕΞΕΙΣ-ΚΛΕΙΔΙΑ

Δομή κοινοτήτων, Κοινότητες, Modularity, Τυχαία δίκτυα, Δίκτυα συσχέτισης, Συντελεστής συσχέτισης, Χρονοσειρές

ABSTRACT

The use of networks is one of the most popular methods when it comes to time series analysis. In this thesis we evaluate the power and size of hypothesis testing, in which the null hypothesis H_0 defines the correlation network as random and without the presence of a community structure. The statistic used to evaluate the hypothesis test is the modularity function. Six different network randomization methods and many different multivariable time series scenarios were used. Finally, the comparison of network randomization methods showed the superiority of the methods that perform randomization in time series compared to those that perform randomization directly on the links of the correlation network.

It is worth mentioning that hypothesis testing also applies to real data. Here it is applied to correlation networks formed by the financial data of Morgan Stanley Capital International that describe the capitalization of 55 nations in the international capital market between March 2004 and March 2009.

KEY WORDS

Community structure, Communities, Modularity, Network randomization, Correlation networks, Correlation coefficient, Time series

Πίνακας Περιεχομένων

Πίνακας Εικόνων

Πίνακας Πινάκων

SUMMARY

In multivariate time series analysis networks whose nodes represent the observed variables and connections are defined in terms of a correlation index for non-directed connections or a causality measure for directed connections, are commonly used. Correlation networks are characterized by a non-random topology due to their transitivity pattern and they display an inherently stronger community structure.

In the survey of networks it is not enough to find the community structure but also to test the statistical significance namely the comparison with the community structure of a set of appropriately constructed random networks. These random networks are derived by rewiring the connections of the initial network so that some characteristics are preserved, such as the number of connections of every node for a binary network or the strength of each node for a weighted network. Another method of generating random networks is through the creation of surrogate random time series and consequently the random network is created by them.

The main purpose of this thesis is the comparison between these two methods of generating random networks and the assessment of their appropriateness for the significance test of modularity regarding the community structure.

In the first chapter the problem of community structure in network analysis is presented. Basic definitions for the community and the partition and the most important applications of community structure are introduced. In the second chapter modularity is reviewed as the most widespread method for community detection, its modifications and the limits it confronts. In the third chapter the randomization methods of networks with main characteristic the randomization of the connections of the initial network are presented thoroughly. In the fourth chapter correlation networks are introduced, and methods for generating random correlation networks through the generation of random covariance matrices and surrogate time series. In the fifth chapter the simulations on benchmark correlation networks and the application in real data are described. Finally in the sixth chapter the conclusions of the thesis are presented.

ΠΡΟΛΟΓΟΣ

Για την ανάλυση πολυμεταβλητών χρονοσειρών είναι σύνηθες η χρησιμοποίηση δικτύων που έχουν ως κόμβους τις παρατηρούμενες μεταβλητές και συνδέσεις που ορίζονται από κάποιο μέτρο συσχέτισης για μη-κατευθυνόμενες συνδέσεις ή αιτιότητας για κατευθυνόμενες συνδέσεις. Έχει παρατηρηθεί ότι τα δίκτυα συσχέτισης χαρακτηρίζονται από μη-τυχαία τοπολογία λόγω του μοτίβου μεταβατικότητας και εμφανίζουν εγγενώς ισχυρότερη δομή κοινοτήτων.

Στην μελέτη των δικτύων δεν αρκεί μόνο να βρεθεί η δομή κοινοτήτων αλλά και να ελεγχθεί η σημαντικότητά της, δηλαδή η σύγκριση της με τη δομή κοινοτήτων ενός συνόλου κατάλληλα κατασκευασμένων τυχαίων δικτύων. Τα τυχαία δίκτυα παράγονται με τυχαιοποίηση των συνδέσεων του αρχικού δικτύου ώστε να διατηρούν κάποια χαρακτηριστικά τους όπως το πλήθος των συνδέσεων κάθε κόμβου για μη-σταθμισμένο δίκτυο ή την ισχύ κάθε κόμβου για σταθμισμένο δίκτυο. Μία άλλη μέθοδος δημιουργίας τυχαίων δικτύων χρησιμοποιεί τη δημιουργία υποκατάστατων-τυχαίων χρονοσειρών και στην συνέχεια το τυχαίο δίκτυο δημιουργείται από αυτές.

Ο κύριος στόχος αυτής της εργασίας είναι η σύγκριση μεταξύ των δύο μεθόδων δημιουργίας τυχαίων δικτύων και η καταλληλόλητά τους στον έλεγχο στατιστικής σημαντικότητας για τη δομή κοινοτήτων.

Στο πρώτο κεφάλαιο παρουσιάζεται το πρόβλημα της δομής κοινοτήτων στην ανάλυση δικτύων. Παρατίθενται βασικοί ορισμοί για την κοινότητα και τη διαμέριση και οι πιο σημαντικές εφαρμογές της δομής κοινοτήτων. Στο δεύτερο κεφάλαιο παρουσιάζεται το μέτρο modularity το πιο γνωστό για την εύρεση κοινοτήτων, οι επεκτάσεις του καθώς και οι περιορισμοί που αντιμετωπίζει. Στο τρίτο κεφάλαιο παρατίθενται διεξοδικά οι μέθοδοι τυχαιοποίησης δικτύων με κύριο χαρακτηριστικό την τυχαιοποίηση των συνδέσεων του αρχικού δικτύου. Στο τέταρτο κεφάλαιο παρουσιάζονται τα δίκτυα συσχέτισης και μέθοδοι δημιουργίας τυχαίων δικτύων συσχέτισης είτε μέσω της δημιουργίας τυχαίου πίνακα συνδιασπορών είτε μέσω δημιουργίας υποκατάστατων χρονοσειρών. Στο πέμπτο κεφάλαιο περιγράφονται οι προσομοιώσεις που έγιναν σε δίκτυα οδηγούς και η εφαρμογή σε πραγματικά δεδομένα. Τέλος στο έκτο κεφάλαιο παρουσιάζονται τα συμπεράσματα της διπλωματικής εργασίας.

ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα διπλωματική εργασία εκπονήθηκε στα πλαίσια του Διατμηματικού Προγράμματος Μεταπτυχιακών Σπουδών «Πολύπλοκα Συστήματα και Δίκτυα».

Θα ήθελα να εκφράσω τις ευχαριστίες μου στον επιβλέποντα καθηγητή μου κ.Κουγιουμτζή Δημήτριο, Καθηγητή της Πολυτεχνικής σχολής του Α.Π.Θ. για την βοήθεια και τις καίριες υποδείξεις που μου παρείχε κατά τη διάρκεια της εκπόνησης της διπλωματικής μου εργασίας.

Τέλος ευχαριστώ την οικογένεια μου για την υποστήριξή τους σε αυτή τη προσπάθειά μου.

Κεφάλαιο 1: Εισαγωγή στη δομή κοινοτήτων.

1.1. Δομή κοινοτήτων.

Η καλύτερα μελετημένη μορφή δομής μεγάλης κλίμακας στα δίκτυα είναι η δομή κοινοτήτων. [1] Τα πραγματικά δίκτυα διαφέρουν αρκετά από τα τυχαία δίκτυα Erdos-Renyi καθώς παρουσιάζουν μεγάλες ανομοιογένειες αποκαλύπτοντας μία υψηλού επιπέδου τάξη και οργάνωση. Η κατανομή του βαθμού των κόμβων είναι πλατιά με την ουρά να ακολουθεί συνήθως κατανομή power law, επομένως πολλοί κόμβοι μικρού βαθμού συνυπάρχουν με λίγους κόμβους μεγάλου βαθμού. Επιπλέον η κατανομή των συνδέσεων δεν είναι μόνο συνολικά αλλά και τοπικά ανομοιογενής, με υψηλή συγκέντρωση συνδέσεων μεταξύ ομάδων κόμβων και χαμηλή συγκέντρωση μεταξύ των ομάδων αυτών. Αυτό το χαρακτηριστικό των πραγματικών δικτύων ονομάζεται δομή κοινοτήτων. Κοινότητες χαρακτηρίζονται ομάδες κόμβων οι οποίοι μοιράζονται μία κοινή ιδιότητα και έχουν παρόμοιο ρόλο μέσα στο δίκτυο. [2]

1.2. Κοινότητες – Διαμέριση του δικτύου

Το πρόβλημα της εύρεσης κοινοτήτων ενώ είναι εύληπτο με την πρώτη ματιά στην πραγματικότητα δεν είναι καλά ορισμένο. Τα κύρια στοιχεία του προβλήματος δηλαδή οι έννοιες της κοινότητας και της διαμέρισης του δικτύου δεν είναι αυστηρώς ορισμένες και απαιτούν ένα βαθμό αυθαιρεσίας και κοινής λογικής. [2]

Πρέπει να σημειωθεί ότι για την εξακρίβωση δομικών κοινοτήτων πρέπει το δίκτυο να είναι αραιό δηλαδή ο αριθμών των συνδέσεων m να είναι της τάξης μεγέθους του αριθμού των κόμβων n του δικτύου. Εάν $m \gg n$ τότε η κατανομή των συνδέσεων μεταξύ των κόμβων γίνεται υπερβολικά ομοιογενής ώστε οι κοινότητες να μην έχουν κάποιο νόημα. [2]

1.2.1. Κοινότητες

Το πρώτο πρόβλημα για την εύρεση κοινοτήτων είναι ένας ποσοτικός ορισμός για την έννοια της κοινότητας. Δεν υπάρχει παγκόσμια αποδεκτός ορισμός. Στην πραγματικότητα ο ορισμός συχνά εξαρτάται από το σύστημα προς μελέτη. Διαισθητικά γίνεται αντιληπτό ότι πρέπει να υπάρχουν περισσότερες συνδέσεις που συνδέουν του κόμβους της κοινότητας μεταξύ τους από ότι συνδέσεις που ενώνουν τους κόμβους της κοινότητας με το υπόλοιπο δίκτυο. Αυτή είναι η βάση για τους περισσότερους ορισμούς της κοινότητας. Επιπλέον στις περισσότερες περιπτώσεις οι κοινότητες

ορίζονται αλγοριθμικά ως αποτέλεσμα κάποιου αλγόριθμου χωρίς την ύπαρξη εκ των προτέρων κάποιου ορισμού. [2]

Έστω ένα υποδίκτυο C ενός δικτύου G, με $|C| = n_C$ και $|G| = n$ κόμβους αντίστοιχα. Ορίζεται ο εσωτερικός και ο εξωτερικός βαθμός κάθε κόμβου $v \in C$, k_v^{int} και k_v^{ext} ως ο αριθμός των συνδέσεων του v με άλλους κόμβους του C ή με το υπόλοιπο δίκτυο αντίστοιχα. Ο εσωτερικός βαθμός k_{int}^C και ο εξωτερικός βαθμός k_{ext}^C του υποδικτύου C είναι το άθροισμα των εσωτερικών βαθμών και αντίστοιχα των εξωτερικών βαθμών των κόμβων του. Ο συνολικός βαθμός k^C του υποδικτύου C είναι εξ ορισμού $k^C = k_{int}^C + k_{ext}^C$.

Ορίζεται $\delta_{int}(C)$ η εσωτερική πυκνότητα του υποδικτύου C ως ο λόγος μεταξύ των εσωτερικών συνδέσεων του C και του αριθμού όλων των πιθανών εσωτερικών συνδέσεων:

$$\delta_{int}(C) = \frac{\left(\sum_{v \in C} k_v^{int}\right)/2}{n_C(n_C - 1)/2} = \frac{\sum_{v \in C} k_v^{int}}{n_C(n_C - 1)}$$

Ομοίως η εξωτερική πυκνότητα $\delta_{ext}(C)$ είναι ο λόγος μεταξύ του αριθμού των συνδέσεων που ενώνουν κόμβους του υποδικτύου C με το υπόλοιπο δίκτυο και τον μέγιστο αριθμό εξωτερικών συνδέσεων:

$$\delta_{ext}(C) = \frac{\sum_{v \in C} k_v^{ext}}{n_C(n - n_C)}$$

Το υποδίκτυο C για να αποτελεί κοινότητα πρέπει η εσωτερική πυκνότητα του $\delta_{int}(C)$ να είναι σημαντικά μεγαλύτερη από την πυκνότητα $\delta(G)$ του δικτύου G και η εξωτερική πυκνότητα $\delta_{ext}(C)$ να είναι αρκετά μικρότερη από την $\delta(G)$. Η εύρεση του καλύτερου συμβιβασμού μεταξύ υψηλής εσωτερικής πυκνότητας $\delta_{int}(C)$ και χαμηλής εσωτερικής πυκνότητας $\delta_{ext}(C)$ είναι έμμεσα ή άμεσα ο στόχος των περισσότερων αλγορίθμων. [2][3]

Μία απαραίτητη ιδιότητα για κάθε κοινότητα είναι η συνδεσιμότητα. Για κάθε υποδίκτυο C πρέπει να υπάρχει ένα μονοπάτι μεταξύ κάθε ζεύγους κόμβων του C που διέρχεται μόνο από κόμβους του υποδικτύου C. Αυτή η ιδιότητα απλοποιεί την διαδικασία εύρεσης κοινοτήτων σε μη-συνδετικά γραφήματα και δίκτυα, αφού σε αυτή τη περίπτωση αναλύονται ξεχωριστά κάθε συνδετικό μέρος του δικτύου. [2]

Οι κοινότητες παρουσιάζουν μεγάλο ενδιαφέρον διότι έχει παρατηρηθεί ότι οι ιδιότητες και τα δομικά χαρακτηριστικά κοινοτήτων του ίδιου δικτύου μπορεί να είναι αρκετά διαφορετικά. Τέτοιες διαφορές μπορεί να είναι σημαντικές για την κατανόηση

της συμπεριφοράς του συστήματος και γίνονται εμφανεί μόνο όταν μελετηθεί η δομή του δικτύου σε επίπεδο κοινοτήτων. [1]

1.2.2. Διαμέριση του δικτύου.

Διαμέριση ενός δικτύου είναι ο διαχωρισμός του σε κοινότητες έτσι ώστε κάθε κόμβος να ανήκει σε μία μόνο κοινότητα. [2]

Ο αριθμός των πιθανών διαμερίσεων σε k κοινότητες ενός δικτύου με n κόμβους είναι ο αριθμός Stirling δεύτερου είδους $S(n,k)$. Ο συνολικός αριθμός όλων των πιθανών διαμερίσεων είναι ο n-στός αριθμός Bell $B_n = \sum_{k=0}^{n} S(n,k)$. [4][5]

Οι διαμερίσεις μπορούν να ταξινομηθούν ιεραρχικά όταν το δίκτυο έχει διαφορετικά επίπεδα οργάνωσης-δομής σε διαφορετικές κλίμακες. [2]

1.3. Σημαντικότητα της δομής κοινοτήτων.

Η δομή κοινοτήτων και οι κοινότητες έχουν σημαντικές εφαρμογές σε πραγματικά δίκτυα. Ομαδοποιώντας τους πελάτες ενός δικτύου διαδικτυακών αγορών με παρόμοια ενδιαφέροντα και μικρή γεωγραφική απόσταση μεταξύ τους καθίσταται δυνατή η βελτίωση της απόδοσης υπηρεσιών που παρέχονται από τον Παγκόσμιο Ιστό εάν κάθε κοινότητα εξυπηρετείται από έναν διακομιστή. [6]

Η εύρεση κοινοτήτων καταναλωτών σύμφωνα με τα αγοραστικά τους ενδιαφέροντα στο διμερές δίκτυο μεταξύ καταναλωτών και προϊόντων ενός διαδικτυακού καταστήματος, παρέχει τη δυνατότητα να χρησιμοποιηθεί ένα αποτελεσματικό σύστημα συστάσεων το οποίο καθοδηγεί καλύτερα τον καταναλωτή στη λίστα των προϊόντων. [7]

Κοινότητες μεγάλων γραφημάτων μπορούν να χρησιμοποιηθούν για να αναπαραστήσουν δομές δεδομένων με σκοπό την αποτελεσματική αποθήκευση των δεδομένων του δικτύου και τη διαχείριση ερωτημάτων πλοήγησης σε αυτές. [8]

Η ανίχνευση κοινοτήτων και των ορίων τους επιτρέπει μία κατηγοριοποίηση των κόμβων σύμφωνα με τη δομική τους θέση στις κοινότητες. Έτσι κόμβοι με κεντρική θέση στις κοινότητες που ανήκουν, δηλαδή έχουν μεγάλο πλήθος συνδέσεων με άλλους κόμβους της ίδιας κοινότητας, έχουν σημαντικό ρόλο για τον έλεγχο και τη σταθερότητα μέσα στην κοινότητα. Κόμβοι οι οποίοι βρίσκονται στα όρια μεταξύ κοινοτήτων παίζουν ένα σημαντικό ρόλο μεσολαβητή και καθοδηγούν τις συνδέσεις

μεταξύ διαφορετικών κοινοτήτων. Αυτή η κατηγοριοποίηση είναι σημαντική σε κοινωνικά [9][10] αλλά και μεταβολικά δίκτυα. [11]

Η εύρεση κοινοτήτων βρίσκει εφαρμογή και στην επιστήμη υπολογιστών. Συγκεκριμένα στον παράλληλο προγραμματισμό είναι σημαντική εύρεση του βέλτιστου τρόπου διανομής εργασιών στους επεξεργαστές ώστε να ελαχιστοποιηθεί η επικοινωνία μεταξύ τους και η απόδοση των υπολογισμών να είναι όσο το δυνατόν γρηγορότερη. Αυτό επιτυγχάνεται με τον διαχωρισμό των επεξεργαστών σε ομάδες περίπου ίδιου πλήθους έτσι ώστε ο αριθμός των φυσικών συνδέσεων μεταξύ επεξεργαστών διαφορετικών κοινοτήτων να είναι όσο το δυνατόν μικρότερος. [2]

Μία άλλη σημαντική διάσταση που σχετίζεται με την δομή κοινοτήτων είναι η ιεραρχική οργάνωση που παρατηρείται στα περισσότερα πραγματικά δίκτυα. Τα πραγματικά δίκτυα αποτελούνται από κοινότητες που περιλαμβάνουν μικρότερες κοινότητες και αυτές με τη σειρά τους περιλαμβάνουν ακόμα μικρότερες κοινότητες. Παραδείγματα ιεραρχικής οργάνωσης είναι το ανθρώπινο σώμα και οι εταιρίες με οργάνωση πυραμίδας. Ο καίριας σημασίας ρόλος της ιεραρχίας έχει αναδειχτεί από τον Simon. [12] Η δημιουργία και η εξέλιξη ενός συστήματος οργανωμένου σε συσχετισμένα μεταξύ τους υποσυστήματα είναι πολύ πιο γρήγορη από ένα μη δομημένο σύστημα, διότι είναι αρκετά πιο εύκολη η συναρμολόγηση των επιμέρους μικρότερων μερών και στην συνέχεια η χρήση τους ως δομικών στοιχείων για μεγαλύτερες δομές μέχρι να ολοκληρωθεί ολόκληρο το σύστημα. [2]

Κεφάλαιο 2: Modularity και εύρεση κοινοτήτων.

2.1.Ποιοτικές συναρτήσεις.

Πολλοί αλγόριθμοι είναι ικανοί να αναγνωρίσουν διαμερίσεις ενός δικτύου σε κοινότητες ωστόσο δεν σημαίνει ότι όλες οι διαμερίσεις που θα βρεθούν είναι το ίδιο καλές. Επομένως είναι βοηθητική, κάποιες φορές και απαραίτητη η ύπαρξη ενός ποιοτικού κριτηρίου για να αξιολογηθεί το πόσο επιτυχημένη είναι η διαμέριση του δικτύου. [2]

Ποιοτική συνάρτηση είναι μία συνάρτηση που αντιστοιχίζει μία τιμή σε κάθε διαμέριση του δικτύου. Με αυτόν τον τρόπο είναι δυνατόν να καταταγούν οι διαμερίσεις με βάση την τιμή που δίνεται από μία ποιοτική συνάρτηση. Διαμερίσεις με υψηλές τιμές θεωρούνται καλές και ή διαμέριση με την υψηλότερη τιμή είναι εξ ορισμού η καλύτερη. [2]

Η ποιοτική συνάρτηση Q είναι προσθετική εάν υπάρχει μία συνάρτηση q τέτοια ώστε για κάθε διαμέριση P ενός δικτύου $Q(P) = \sum_{C \in P} q(C)$ όπου C είναι μία κοινότητα της διαμέρισης P. Με αυτόν τον τρόπο η ποιότητα μίας διαμέρισης δίνεται από το άθροισμα των ιδιοτήτων κάθε μιας ξεχωριστής κοινότητας. Οι περισσότερες ποιοτικές συναρτήσεις στη βιβλιογραφία είναι προσθετικές, ωστόσο δεν είναι μία απαραίτητη προϋπόθεση. [2]

2.2.Modularity

Η πιο γνωστή ποιοτική συνάρτηση είναι η συνάρτηση modularity των Newman και Girvan. Βασίζεται στην ιδέα ότι ένα τυχαίο δίκτυο δεν αναμένεται να έχει δομή κοινοτήτων, έτσι η πιθανή ύπαρξη κοινοτήτων αποκαλύπτεται από τη σύγκριση μεταξύ της πραγματικής πυκνότητας των ακμών σε ένα υπό-δίκτυο και της αναμενόμενης πυκνότητας στο ίδιο υπό-δίκτυο εάν οι συνδέσεις είχαν τοποθετηθεί τυχαία. [2]

Η αναμενόμενη πυκνότητα ακμών εξαρτάται από το επιλεγόμενο μοντέλο τυχαίου δικτύου (null model) δηλαδή ένα αντίγραφο του πραγματικού δικτύου που κρατάει κάποιες δομικές ιδιότητες αλλά χωρίς δομή κοινοτήτων. [13]

Η συνάρτηση modularity γράφεται ως εξής:

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - P_{ij}) \delta(C_i, C_j)$$

όπου το άθροισμα ελέγχει όλα τα ζευγάρια των κόμβων, A είναι ο πίνακας γειτνίασης του μη κατευθυνόμενου δικτύου, m το σύνολο όλων των συνδέσεων του δικτύου, και P_{ij} αντιπροσωπεύει τον αναμενόμενο αριθμό ακμών μεταξύ των κόμβων i και j στο null

model. Η συνάρτηση δ δίνει 1 αν οι κόμβοι i και j ανήκουν στην ίδια κοινότητα ($C_i = C_j$), αλλιώς 0. [13]

Η επιλογή του null model μπορεί να γίνει αυθαίρετα και υπάρχουν αρκετές πιθανές επιλογές. Ωστόσο τα πραγματικά δίκτυα χαρακτηρίζονται από αρκετά πλατιές κατανομές βαθμού και λόγω των σημαντικών συνεπειών που έχουν αυτές στην λειτουργία και στη δομή των πραγματικών δικτύων είναι προτιμότερο να χρησιμοποιηθεί ένα null model που θα διατηρεί με την ίδια κατανομή βαθμού με το πραγματικό δίκτυο. [13]

Το τυπικό null model για την συνάρτηση modularity ονομάζεται και Newman-Girvan null model και επιβάλλει ότι η αναμενόμενη βαθμική ακολουθία θα είναι ίδια με την πραγματική βαθμική ακολουθία του δικτύου και στην ουσία είναι ισοδύναμο με το configuration model (παράγραφος 2.7). [2] Έστω ένα μη κατευθυνόμενο δίκτυο, βαθμική ακολουθία ονομάζεται η μονοτονική μη αύξουσα ακολουθία των βαθμών των κόμβων του δικτύου. Το άθροισμα των στοιχείων οποιασδήποτε βαθμικής ακολουθίας είναι πάντα άρτιος αριθμός.

Σε αυτό το null model ένας κόμβος μπορεί να συνδεθεί με οποιοδήποτε άλλο κόμβο του δικτύου και η πιθανότητα ότι οι κόμβοι i και j με βαθμούς k_i και k_j συνδέονται μπορεί να υπολογιστεί χωρίς πρόβλημα. Για να δημιουργηθεί μία σύνδεση μεταξύ των i και j πρέπει να ενωθούν 2 μισές ακμές (stubs) προσπίπτουσες στους κόμβους αυτούς. Η πιθανότητα p_i να επιλεγεί τυχαία ένα stub από τον κόμβο i είναι $k_i/2m$ αφού υπάρχουν k_i stubs που ακουμπούν στον κόμβο i και συνολικά υπάρχουν 2m. Η πιθανότητα για να δημιουργηθεί μία σύνδεση μεταξύ των i και j είναι $p_i p_j = \frac{k_i k_j}{4m^2}$ αφού οι συνδέσεις τοποθετούνται ανεξάρτητα η μία από την άλλη. Επομένως ο αναμενόμενος αριθμός συνδέσεων είναι $P_{ij} = 2m p_i p_j = \frac{k_i k_j}{2m}$ για τους κόμβους i και j. Άρα η τελική έκφραση για το modularity είναι:

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(C_i, C_j).$$

Εφόσον οι συνεισφορές του αθροίσματος προέρχονται από κόμβους οι οποίοι ανήκουν στην ίδια κοινότητα ο τύπος για το modularity μπορεί να γραφεί ως άθροισμα στις κοινότητες ως εξής:

$$Q = \sum_{c=1}^{n_c}\left[\frac{l_c}{m} - \left(\frac{d_c}{2m}\right)^2\right]$$

όπου n_c ο αριθμός των κοινοτήτων, l_c ο συνολικός αριθμός των συνδέσεων που ενώνουν κόμβους της κοινότητας c και d_c το άθροισμα των βαθμών των κόμβων της κοινότητας c. Ο πρώτος όρος κάθε άθροισης είναι το κλάσμα των εσωτερικών συνδέσεων της κοινότητας ενώ ο δεύτερος όρος αντιπροσωπεύει το κλάσμα των αναμενόμενων συνδέσεων που θα υπήρχαν στο τυχαίο δίκτυο με τον ίδιο αναμενόμενο βαθμό για κάθε κόμβο. [2]

Σύμφωνα με αυτή τη μορφή του modularity ένα υπό-δίκτυο μπορεί να θεωρηθεί ως κοινότητα αν η συνεισφορά του στο άθροισμα είναι θετική. Όσο περισσότερο ο αριθμός των εσωτερικών συνδέσεων μιας κοινότητας υπερβαίνει τον αναμενόμενο αριθμό, τόσο καλύτερα ορισμένη είναι η κοινότητα. Έτσι μεγάλες θετικές τιμές του modularity υποδεικνύουν καλές διαμερίσεις. [2]

Η μέγιστη τιμή του modularity ενός δικτύου μεγαλώνει αν το μέγεθος του δικτύου αυξηθεί, επομένως το modularity δεν πρέπει να χρησιμοποιείται για την σύγκριση της ποιότητας της δομής κοινοτήτων μεταξύ δικτύων που διαφέρουν σε μέγεθος. [2]

Το modularity παίρνει τιμές στο διάστημα [-1,1]. Στην περίπτωση που θεωρηθεί ολόκληρο το δίκτυο ως μία κοινότητα το modularity είναι 0 αφού οι δύο όροι του αθροίσματος είναι ίσοι. Επίσης για την διαμέριση στην οποία κάθε κόμβος θεωρείται ως ξεχωριστή κοινότητα παίρνει πάντα αρνητική τιμή. Σε αυτή την περίπτωση το άθροισμα περιλαμβάνει n όρους (όσοι και οι κόμβοι του δικτύου) οι οποίοι είναι όλοι αρνητικοί αφού οι πρώτοι όροι είναι μηδέν. [2]

2.3. Τροποποιήσεις του modularity.

Στην σχετική βιβλιογραφία με την εύρεση κοινοτήτων υπάρχουν αρκετές μετατροπές και επεκτάσεις του modularity οι οποίες πηγάζουν από τις ειδικές κλάσεις γραφημάτων που υπάρχουν.

Το modularity έχει επεκταθεί και για σταθμισμένα δίκτυα. [14] Πρέπει να αντικατασταθούν οι βαθμοί k_i και k_j με την ισχύ s_i και s_j των κόμβων i και j αντίστοιχα. Για κατάλληλη κανονικοποίηση ο αριθμός των συνδέσεων m πρέπει να αντικατασταθεί από το άθροισμα W των βαρών όλων των συνδέσεων.

Είναι το αναμενόμενο βάρος της σύνδεσης μεταξύ των κόμβων i και j στο null model για το modularity και πρέπει να συγκριθεί με το πραγματικό βάρος W_{ij} της σύνδεσης στο πραγματικό δίκτυο.

Αυτό μπορεί να γίνει εύκολα κατανοητό για την περίπτωση όπου όλα τα βάρη είναι πολλαπλάσια μίας μονάδας βάρους ώστε να μπορούν να γραφούν ως ακέραια πολλαπλάσια του. Επομένως το βάρος μιας σύνδεσης μπορεί να αντικατασταθεί από τόσες συνδέσεις όσο είναι το ακέραιο πολλαπλάσιο του βάρους σε μονάδες βάρους. Για το παραγόμενο multigraph μπορεί να ακολουθηθεί η ίδια διαδικασία όπως στα μη σταθμισμένα δίκτυα η οποία δίνει ως αποτέλεσμα

$$Q_w = \frac{1}{2W}\sum_{ij}\left(W_{ij} - \frac{s_i s_j}{2W}\right)\delta(C_i C_j).$$

Αντίστοιχα η έκφραση για το modularity σε σταθμισμένα δίκτυα μπορεί να γραφεί και ως άθροισμα στις κοινότητες:

$$Q_w = \sum_{c=1}^{n_c}\left[\frac{W_c}{W} - \left(\frac{S_c}{2W}\right)^2\right]$$

Όπου W_c είναι το άθροισμα των βαρών των εσωτερικών συνδέσεων της κοινότητας c και S_c το άθροισμα των βαρών όλων των κόμβων που ανήκουν στην κοινότητα c. [2]

Το modularity έχει επεκταθεί και σε κατευθυνόμενα δίκτυα [15] αλλά και σε δίκτυα με αρνητικά βάρη. [16]

2.4. Μοντέλα τυχαίων δικτύων για την μεγιστοποίηση του modularity.

Η διαδικασία εύρεσης κοινοτήτων μεγιστοποιώντας τη συνάρτηση modularity εξαρτάται εξ ορισμού από την επιλογή του null model. Η επιλογή ενός null model είναι σημαντική διότι μπορεί να έχει σημαντική επίδραση στην δομή κοινοτήτων που θα αποκτήσουμε μέσω της βελτιστοποίησης μίας ποιοτικής συνάρτησης και γιατί αλλάζει την ερμηνεία των κοινοτήτων. Επομένως είναι χρήσιμο να επιλεχθεί ένα null model που να παίρνει υπόψιν του κατάλληλα χαρακτηριστικά του υπό μελέτης συστήματος. [17] Η επιλογή του null model είναι αυθαίρετη και υπάρχουν πολλές επιλογές πέρα από το τυπικό null model των Newman Girvan το οποίο πολλές φορές θεωρείται ως μη ρεαλιστικό με την έννοια ότι κάθε κόμβος μπορεί να συνδεθεί με οποιοδήποτε άλλο κόμβο χωρίς προτίμηση . [2][18]

2.4.1. Uniform null model.

Το uniform null model είναι το πιο απλό που συναντάται στη βιβλιογραφία. Οι δύο μοναδικές του προϋποθέσεις είναι ότι το δίκτυο κρατάει τον ίδιο αριθμό συνδέσεων με το πραγματικό δίκτυο και ότι οι συνδέσεις τοποθετούνται με την ίδια

πιθανότητα μεταξύ οποιοδήποτε κόμβων. Ο όρος του null model θα είναι σταθερός για όλα τα ζευγάρια κόμβων:

$$P_{ij} = p = \frac{2m}{n(n-1)}, \forall i,j$$

Επομένως παράγεται ένα Bernoulli τυχαίο δίκτυο αφού η παρουσία και η απουσία μίας σύνδεσης είναι ανεξάρτητη και ισόνομη τυχαία μεταβλητή Bernoulli με πιθανότητα p.

Ωστόσο το uniform null model δεν περιγράφει καλά τα πραγματικά δίκτυα διότι έχει κατανομή βαθμού Poisson η οποία είναι διαφορετική από τις κατανομές βαθμού που συναντώνται στα πραγματικά δίκτυα. [2]

2.4.2. Distance null model.

Το distance null model λαμβάνει υπόψιν του την έλξη ομοιότητας (similarity attraction) δηλαδή ότι οι συνδέσεις τείνουν να ενώνουν κόμβους οι οποίοι είναι όμοιοι μεταξύ τους και είναι κατάλληλο για δίκτυα των οποίων οι κόμβοι περιλαμβάνουν επιπλέον πληροφορίες που θα χρησιμοποιηθούν για την εκτίμηση της ομοιότητας των κόμβων. [18]

Έστω d_{ij} η απόσταση ομοιότητας μεταξύ των κόμβων i και j (όσο μικρότερη είναι η τιμή d_{ij} τόσο πιο όμοιοι θεωρούνται οι δύο κόμβοι) μπορεί να εκτιμηθεί από μία συνάρτηση απόστασης η οποία λαμβάνει ως εισακτέα τιμή διαθέσιμες πληροφορίες για τους κόμβους i και j όπως για παράδειγμα πληροφορίες για τη δομή του δικτύου. Γενικά πρέπει να ικανοποιεί του ακόλουθους περιορισμούς:

1. $d_{ij} \geq 0$ η ισότητα ισχύει αν i=j
2. $d_{ij} = d_{ji}$
3. $d_{ij} \leq d_{it} + d_{tj}$

Οι συνδέσεις τοποθετούνται τυχαία και το αναμενόμενο πλήθος συνδέσεων μεταξύ των κόμβων i και j είναι:

$$P_{ij}^{DIST} = (\widetilde{P_{ij}} + \widetilde{P_{ji}})/2$$

Και

$$\widetilde{P_{ij}} = \frac{k_i k_j e^{-(d_{ij}/\sigma)^2}}{\sum_{t=1}^{n} k_t e^{-(d_{ti}/\sigma)^2}}$$

Όπου $\sigma \in (0, +\infty)$.

Η δράση στον κόμβο j από το πεδίο του κόμβου i υπολογίζεται $\varphi_i(j) = k_i e^{-(d_{ij}/\sigma)^2}$ όπου η δύναμη του κόμβου i είναι ο βαθμός του k_i και η συνάρτηση $f(d) = e^{-(d/\sigma)^2}$ παίρνει τιμές στο διάστημα (0,1] μειώνεται μονοτονικά με το d. Η παράμετρος $\sigma \in (0, +\infty)$ ρυθμίζει το εύρος αλληλεπίδρασης του πεδίου, εάν σ έχει μεγάλη τιμή η $f(d)$ μειώνεται γρήγορα υποδεικνύοντας ένα πεδίο με μεγάλη ακτίνα δράσης.

Σημειώνεται ότι τα πεδία που δημιουργούνται από διαφορετικούς κόμβους μπορεί να επικαλύπτονται. Επομένως το επικαλυπτόμενο πεδίο στον κόμβο j είναι:

$$\varphi_{sup}(j) = \sum_{t=1}^{n} \varphi_t(j) = \sum_{t=1}^{n} k_i e^{-(d_{tj}/\sigma)^2}$$

Και προκύπτει ότι το $\widetilde{P_{ij}}$ είναι αποτέλεσμα της δράσης στον κόμβο i των πεδίων που δημιουργούνται ξεχωριστά από τον i και τον j διαιρούμενο με την επικαλυπτόμενη δράση των πεδίων όλων των κόμβων στον κόμβο i.

$$\widetilde{P_{ij}} = \frac{\varphi_i(i)\varphi_j(i)}{\varphi_{\sup(i)}}$$

Το distance null model έχει κατά προσέγγιση την ίδια βαθμική κατανομή με το πραγματικό δίκτυο. Επίσης τα στοιχεία $\widetilde{P_{ij}}$ είναι θετικά συσχετισμένα με τον βαθμό k_i και αρνητικά συσχετισμένα με την απόσταση d_{ij}, δηλαδή οι συνδέσεις τείνουν να ενώνουν κόμβου υψηλού βαθμού που είναι όμοιοι μεταξύ τους.

Στην ακραία περίπτωση για $\sigma \to 0+$ η ακτίνα δράσης του πεδίου που δημιουργείται από τους κόμβους είναι τόσο μικρή που το πεδίο δρα μόνο με τον κόμβο που το δημιουργεί. Σαν αποτέλεσμα κάθε κόμβος μπορεί να δράσει μόνο με τον εαυτό του.

Ενώ στην άλλη ακραία περίπτωση για $\sigma \to +\infty$ η ακτίνα δράσης του πεδίου που δημιουργείται από κάθε κόμβο γίνεται τόσο μεγάλη ώστε αλληλοεπιδρά με όλους τους κόμβους το ίδιο. Η έλξη ομοιότητας εξαφανίζεται και το distance null model γίνεται ισοδύναμο με το Newman-Girvan null model. [18]

2.4.3. Null model μερικώς δεσμευμένο από τους βαθμούς των κόμβων.

Το συγκεκριμένο null model δημιουργήθηκε από τους Chang, Leahy και Pantazis και υπολογίζει το αναμενόμενο πλήθος ή το αναμενόμενο βάρος (για μη-σταθμισμένο ή σταθμισμένο δίκτυο αντίστοιχα) μίας σύνδεσης δεδομένου του βαθμού των δύο κόμβων στους όποιους προσπίπτει. [19]

Έστω $E(A_{ij} \mid k_i, k_j)$ το αναμενόμενο βάρος της σύνδεσης μεταξύ των κόμβων i και j δεδομένου των βαθμών τους k_i και k_j. Η ιδέα βασίζεται στην παρατήρηση ότι η σημαντικότητα μιας σύνδεσης είναι απευθείας συνδεδεμένη με το σύνολο των συνδέσεων των κόμβων της. Εάν δύο κόμβοι έχουν υψηλούς βαθμούς υπάρχει μεγάλη πιθανότητα να είναι συνδεδεμένοι ακόμα και σε ένα τυχαίο δίκτυο που δεν παρουσιάζει δομή κοινοτήτων. Το αντίθετο ισχύει για τους κόμβους με χαμηλό βαθμό, όπου ακόμα και ασθενής συνδέσεις μπορεί να είναι σημαντικές.

Ενώ το Newman-Girvan null model έχει εξάρτηση από τους βαθμούς των κόμβων δεν είναι ξεκάθαρη μορφή μίας δεσμευμένης αναμενόμενης τιμής. Η δεσμευμένη αναμενόμενη τιμή για την σύνδεση A_{ij} υπολογίζεται χρησιμοποιώντας το θεώρημα Bayes:

$$E\big(A_{ij} \mid k_i, k_j\big) = \int t \cdot P\big(A_{ij} = t \mid k_i, k_j\big) dt$$

$$= \int t \cdot \frac{P\big(k_i, k_j \mid A_{ij} = t\big) P\big(A_{ij} = t\big)}{\int P\big(k_i, k_j \mid A_{ij} = u\big) P\big(A_{ij} = u\big) du} dt$$

Για να επιλυθεί αυτή η εξίσωση πρέπει να οριστεί η από κοινού κατανομή των συνδέσεων του δικτύου.

Για σταθμισμένο δίκτυο ελέγχεται η περίπτωση Gaussian τυχαίου δικτύου N κόμβων με ανεξάρτητες και ισόνομες συνδέσεις με μέση τιμή μ και διασπορά σ^2. Το αναμενόμενο βάρος της σύνδεσης μεταξύ i και j είναι:

$$E\big(A_{ij} \mid k_i, k_j\big) = \begin{cases} \frac{k_i + k_j - (N-2)\mu}{N}, & i \neq j \\ 0, & i = j \end{cases} .$$

Για μη σταθμισμένο δίκτυο ελέγχεται η περίπτωση όπου όλες οι συνδέσεις είναι ανεξάρτητες και ισόνομες τυχαίας μεταβλητές που ακολουθούν κατανομή Bernoulli με παράμετρο p, όπου p η πιθανότητα ύπαρξης σύνδεσης:

$$E\big(A_{ij} \mid k_i, k_j\big) = \begin{cases} \dfrac{k_i k_j}{k_i k_j + (N-1-k_i)(N-1-k_j) \cdot p/(1-p)}, & i \neq j \\ 0, & i = j \end{cases}$$

2.4.4. Null model πλήρως δεσμευμένο από τους βαθμούς των κόμβων.

Το προηγούμενο null model επεκτείνεται θεωρώντας το αναμενόμενο βάρος της σύνδεσης μεταξύ των κόμβων i και j δεδομένου των βαθμών όλων των κόμβων του δικτύου. [19]

$$E\left(A_{ij} \mid k_1, k_2, \dots, k_N\right) = E\left(A_{ij} \mid \mathrm{k}\right)$$

Αρχικά μετασχηματίζονται τα στοιχεία του πίνακα γειτνίασης A σε ένα διάνυσμα στήλη x χρησιμοποιώντας το μετασχηματισμό $A_{ij} = x_l$ όπου $l = \frac{(2N-j)(j-1)}{2} + (i-j), \forall (0 < j < i < N)$. Αυτός ο μετασχηματισμός λαμβάνει υπόψιν του τη συμμετρική φύση του πίνακα A και χρησιμοποιεί μόνο τα στοιχεία του κάτω τριγωνικού μέρους του πίνακα αποκλείοντας τα στοιχεία της διαγώνιου.

Θεωρείται η γραμμική σχέση $Hx = k$, όπου H είναι ο $N \times L$ πίνακας προσπτώσεων του δικτύου που ενώνει τους N κόμβους με τις L συνδέσεις του δικτύου. Επομένως το αναμενόμενο βάρος της σύνδεσης μεταξύ των κόμβων i και j δεδομένου των βαθμών όλων των κόμβων του δικτύου:

$$E(x \mid \mathrm{k}) = E(x \mid \mathrm{Hx} = \mathrm{k}) = \int x \cdot P(x \mid \mathrm{Hx} = \mathrm{k}) dx$$

Δεν υπάρχει αναλυτική λύση για την δεσμευμένη πιθανότητα $P(x \mid \mathrm{Hx} = \mathrm{k})$, μπορεί να υπολογιστεί για πεπερασμένο πλήθος για την πολυμεταβλητή κανονική κατανομή. Δηλαδή για το τυχαίο δίκτυο του οποίου οι συνδέσεις ακολουθούν κανονική κατανομή με διάνυσμα μέσων τιμών μ_x και πίνακα συνδιασπορών Σ_x. Δεδομένου ότι το διάνυσμα k είναι γραμμικός συνδυασμός του x ακολουθεί πολυμεταβλητή κανονική κατανομή με μέση τιμή $\mu_k = H\mu_x$ και πίνακα συνδιασπορών $\Sigma_k = H\Sigma_x H^T$. Η δεσμευμένη πιθανότητα $P(x \mid \mathrm{Hx} = \mathrm{k})$ ακολουθεί πολυμεταβλητή κανονική κατανομή με:

$$E(x \mid \mathrm{Hx} = \mathrm{k}) = \mu_{x|k} = \mu_x + \Sigma_{xk}\Sigma_k^{-1}(k - \mu_k)$$

$$\Sigma_{x|k} = \Sigma_x - \Sigma_{xk}\Sigma_k^{-1}\Sigma_{kx}$$

Όπου $\Sigma_{xk} = \Sigma_x H^T$.

Για την ειδική περίπτωση ενός Gaussian τυχαίου δικτύου με ανεξάρτητες και ισόνομες συνδέσεις με $\mu_x = \mu 1$ και $\Sigma_x = \sigma^2 I$ η δεσμευμένη αναμενόμενη τιμή του l στοιχείου του διανύσματος x ή αλλιώς ισοδύναμα το A_{ij} στοιχείο του πίνακα A είναι:

$$E\left(A_{ij} \mid \mathrm{k}\right) = E(x_l \mid \mathrm{k}) = \begin{cases} \dfrac{k_i + k_j}{N-2} - \dfrac{2m}{(N-1)(N-2)}, & i \neq j \\ 0, & i = j \end{cases}$$

Όπου m η συνολική ισχύς του δικτύου. Ο πρώτος όρος στο δεξί μέρος της παραπάνω εξίσωσης για $i \neq j$ δείχνει ότι η αναμενόμενη τιμή μίας συγκεκριμένης σύνδεσης είναι θετικά συσχετισμένη με την ισχύ των 2 κόμβων στους οποίους προσπίπτει, ενώ ο δεύτερος όρος δείχνει ότι είναι μειώνεται όταν η συνολική ισχύς του δικτύου μεγαλώνει ενώ η ισχύς των δύο κόμβων μένει σταθερή. [19]

2.4.5. Null model χωρικών δικτύων.

Με την αυξανόμενη ποσότητα δεδομένων εξαρτημένων από τον χώρο είναι απαραίτητο να αναπτυχθούν τεχνικές για την εύρεση κοινοτήτων οι οποίες εκμεταλλεύονται την επιπρόσθετη χωρική πληροφορία. Επιπλέον η δομή των δικτύων επηρεάζεται ριζικά από χωρικές επιδράσεις. [17]

Χωρικό δίκτυο (spatial networks, geometric graph) είναι ένα δίκτυο στο οποίο οι κόμβοι και οι συνδέσεις είναι χωρικά στοιχεία συσχετισμένα με γεωμετρικά αντικείμενα, δηλαδή οι κόμβοι είναι τοποθετημένοι σε έναν χώρο εφοδιασμένο με μία μετρική. Το Internet, δίκτυα κινητής τηλεφωνίας, νευρωνικά δίκτυα, κοινωνικά δίκτυα είναι παραδείγματα όπου ο υποβόσκων χώρος είναι συναφής και όπου η τοπολογία του δικτύου δεν περιλαμβάνει εξ ολοκλήρου την πληροφορία. [20] Τα συγκεκριμένα null model αναφέρονται σε σταθμισμένα δίκτυα.

2.4.5.1. Gravity null model.

Σε πολλά χωρικά δίκτυα η εγγύτητα έχει ισχυρή επίδραση στις συνδέσεις μεταξύ των κόμβων, δηλαδή γειτονικοί κόμβοι είναι πιο πιθανό να συνδεθούν μεταξύ τους (και οι συνδέσεις τους είναι πιο πιθανό να έχουν μεγαλύτερα βάρη) από ότι κόμβοι που είναι απομακρυσμένοι.

Το gravity null model προτάθηκε από τους Expert και Blondel [21] και υποθέτει ότι η αλληλεπίδραση μεταξύ δύο τοποθεσιών-κόμβων είναι αναλογική με την σημαντικότητά τους (πληθυσμός) και φθίνει με την απόστασή τους. [17]

Στο τυπικό gravity model η αλληλεπίδραση μεταξύ των κόμβων i και j με αντιστοιχούν πληθυσμούς n_i και n_j και με την μεταξύ τους απόσταση να είναι d_{ij} είναι:

$$G_{ij} = n_i^{\alpha} n_j^{\beta} f(d_{ij})$$

Όπου η συνάρτηση $f(d)$ περιγράφει την επίδραση του χώρου στις αλληλεπιδράσεις των κόμβων. Συνήθεις επιλογές για την συνάρτηση είναι: $f(d_{ij}) = 1/d_{ij}$, $f(d_{ij}) =$

$1/d_{ij}^2$, $f(d_{ij}) = e^{-d_{ij}}$ και $f(d_{ij}) = d_{ij}^{\kappa}$ Οι παράμετροι α,β και κ εκτιμώνται με παλινδρόμηση.

Η πιο απλή μορφή του δίνεται για $\alpha = \beta = 1$ και $\kappa = -1$ και το gravity null model γίνεται:

$$P_{ij}^{grav} = I_i I_j f(d_{ij})$$

Η συνάρτηση $f(d)$ εκτιμάται για κάθε απόσταση d για όλα τα ζεύγη κόμβων που απέχουν απόσταση d ως εξής:

$$f(d) = \frac{\sum_{\{k,l|d_{kl}=d\}} W_{kl}}{\sum_{\{k,l|d_{kl}=d\}} (I_k I_l)}$$

Ως μέτρο της σημαντικότητας κάθε κόμβου I_i μπορεί να χρησιμοποιηθεί ο πληθυσμός n_i, η πυκνότητα πληθυσμού ή ο λογάριθμος του πληθυσμού χωρίς να έχουν ποιοτικές διαφορές στην δομή κοινοτήτων. Μία άλλη απλή επιλογή για την σημαντικότητα κάθε κόμβου είναι να χρησιμοποιηθεί η ισχύς κάθε κόμβου, ωστόσο το null model γίνεται αρκετά όμοιο με το Newman-Girvan null model.

Επομένως η τελική μορφή του gravity null model είναι:

$$P_{ij}^{grav} = I_i I_j \frac{\sum_{\{k,l|d_{kl}=d\}} W_{kl}}{\sum_{\{k,l|d_{kl}=d_{ij}\}} (I_k I_l)}$$

2.4.5.2. Radiation null model.

Το gravity null model περιλαμβάνει πολλές παραμέτρους που πρέπει να εκτιμηθούν από τα δεδομένα ή να επιλεγούν αυθαίρετα. Επίσης λόγω κατασκευής του δεν μπορεί να προβλέψει διαφορετικές ροές μεταξύ τοποθεσιών που έχουν την ίδια απόσταση αλλά έχουν περιοχές με διαφορετικές πυκνότητες πληθυσμού ενδιάμεσά τους. [17]

Το radiation null model έχει σχεδιαστεί για να αντιμετωπίσει αυτά τα προβλήματα. [22] Η μέση μετακινούμενη ροή που προβλέπεται από το radiation model για δύο τοποθεσίες i και j με πληθυσμούς n_i και n_j αντίστοιχα είναι:

$$T_{ij} = T_i \frac{n_i n_j}{(n_i + r_{ij})(n_i + n_j + r_{ij})}$$

Όπου r_{ij} είναι ο πληθυσμός μεταξύ των τοποθεσιών i και j, και T_i είναι ο αριθμός των φορέων στην τοποθεσία i.

Για να αποφευχθεί η δημιουργία κατευθυνόμενου δικτύου χρησιμοποιείται η συμμετρική προβλεπόμενη ροή

$$\hat{T}_{ij} = (T_{ij} + T_{ji})/2$$

Χρησιμοποιώντας την ίδια λογική με το gravitation null model για την κατασκευή του radiation null model:

$$P_{ij}^{rad} = \hat{T}_{ij} \frac{\sum_{\{k,l|d_{kl}=d\}} W_{kl}}{\sum_{\{k,l|d_{kl}=d_{ij}\}} \hat{T}_{kl}}$$

2.4.6. Null model δικτύων συσχέτισης.

Η ιδέα χρησιμοποίησης αλγορίθμων ευρέσεων κοινοτήτων για την ανάλυση χρονοσειρών στην ουσία αντικαθιστά δεδομένα δικτύου με πίνακες διασυσχέτισης. Η προσέγγιση αυτή υποφέρει από τον περιορισμό ότι το null model για το modularity είναι ασύμφωνο με τις ιδιότητες των πινάκων συσχέτισης. [23] Αντιμετωπίζοντας ένα εμπειρικό πίνακα συσχέτισης C διάστασης $N \times N$ ως ένα σταθμισμένο δίκτυο το modularity ορίζεται ως εξής:

$$Q = \frac{1}{2C_{norm}} \sum_{ij} (C_{ij} - P_{ij}) \delta(\sigma_i, \sigma_j)$$

Όπου $C_{norm} = \sum_{ij} C_{ij}$, $P_{ij} = \frac{k_i k_j}{C_{norm}}$ και $k_i = \sum_j C_{ij}$.

Το πρόβλημα προκύπτει λόγω του ορισμού του null model το οποίο είναι έγκυρο για έναν πίνακα γειτνίασης A αλλά όχι αν αυτός αντικατασταθεί από έναν πίνακα συσχέτισης C.

Από τη θεωρία τυχαίων πινάκων (RMT) ένας πίνακας συσχέτισης που δημιουργείται από N χρονοσειρές με T παρατηρήσεις για κάθε μία (στα όρια $N \to +\infty$ και $T \to +\infty$ με $1 < T/N < +\infty$) έχει κατανομή ιδιοτιμών γνωστή ως Marcenko-Pastur ή Sengupta-Mistra:

$$\rho(\lambda) = \frac{T}{N} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\lambda} \, , \lambda_- < \lambda < \lambda_+$$

και $\rho(\lambda) = 0$, αλλιώς. Όπου η μέγιστη λ_+και η ελάχιστη λ_- ιδιοτιμή δίνονται από τον τύπο:

$$\lambda_\pm = \left(1 \pm \sqrt{\frac{N}{T}}\right)^2$$

Ο κύριος όγκος των ιδιοτιμών ενός εμπειρικού πίνακα συσχέτισης βρίσκεται μέσα στο διάστημα $[\lambda_-, \lambda_+]$ και μπορεί να θεωρηθεί ως αποτέλεσμα τυχαίου θορύβου. Όσες ιδιοτιμές είναι μεγαλύτερες από την μέγιστη ιδιοτιμή λ_+ που προβλέπει η

κατανομή Marcenko-Pastur θεωρούνται ότι αντιπροσωπεύουν σημαντική δομή στα δεδομένα. Επομένως κάθε εμπειρικός πίνακας συσχέτισης C μπορεί να γραφεί ως άθροισμα δύο πινάκων:

$$C = C^{(r)} + C^{(s)}$$

όπου $C^{(r)}$ είναι το τυχαίο τμήμα που απαρτίζεται από τις ιδιοτιμές λ_i που είναι μικρότερες ή ίσες από τη λ_+ (συνήθως συμπεριλαμβάνονται και οι ιδιοτιμές που είναι μικρότερες από την λ_-) και τα αντίστοιχα ιδιοδιανύσματα και $C^{(s)}$ είναι το τμήμα που αποτελείται από τις υπόλοιπες ιδιοτιμές που είναι μεγαλύτερες από τη λ_+.

Ένα σημαντικό χαρακτηριστικό των εμπειρικών πινάκων συσχέτισης είναι ότι η μεγαλύτερη ιδιοτιμή λ_m είναι αρκετά μεγαλύτερη από τις υπόλοιπες, το αντίστοιχο ιδιοδιάνυσμα έχει μόνο θετικά πρόσημα και μπορεί να αναγνωριστεί ως η «καθολική λειτουργία» (market mode ή global mode) δηλαδή ένας κοινός παράγοντας που επηρεάζει όλες τις μεταβλητές του συστήματος. Επομένως ο πίνακας C μπορεί να γραφεί τώρα ως:

$$C = C^{(r)} + C^{(g)} + C^{(m)}$$

όπου $C^{(m)}$ το τμήμα που ορίζεται από την μεγαλύτερη ιδιοτιμή λ_m και το αντίστοιχο ιδιοδιάνυσμα και $C^{(g)}$ το τμήμα που αποτελείται από της υπόλοιπες ιδιοτιμές που είναι μεγαλύτερες της λ_+. Σαν αποτέλεσμα ορίζονται τρία null model για πίνακες συσχέτισης. [23]

2.4.6.1. Χρονοσειρές χωρίς τάση – άπειρες παρατηρήσεις.

Στην περίπτωση χρονοσειρών άπειρων παρατηρήσεων χωρίς τάση για να ανιχνευθούν κοινότητες χρονοσειρών που είναι περισσότερο συσχετισμένες από το αναμενόμενο και υπό την μηδενική υπόθεση ότι όλες οι χρονοσειρές είναι ανεξάρτητες μεταξύ τους ανά δύο, το null model ορίζεται ως ο ταυτοτικός πίνακας I διάστασης $N \times N$. [23]

$$P_{ij} = I_{ij}$$

2.4.6.2. Χρονοσειρές χωρίς τάση – πεπερασμένο πλήθος παρατηρήσεων.

Για χρονοσειρές με πεπερασμένο πλήθος παρατηρήσεων και χωρίς τάση το κατάλληλο null model ώστε να αναμένει μία ποσότητα τυχαίου θόρυβου [23] είναι:

$$P_{ij} = C_{ij}^{(r)}$$

2.4.6.3. Χρονοσειρές με τάση – πεπερασμένο πλήθος παρατηρήσεων.

Στην περίπτωση χρονοσειρών με τάση και πεπερασμένο πλήθος παρατηρήσεων το κυρίαρχο τμήμα $C^{(m)}$ του πίνακα συσχέτισης C θα μεγιστοποιήσει το modularity για την τετριμμένη διαμέριση όπου όλες οι χρονοσειρές θα ανήκουν στη ίδια κοινότητα. Επομένως για να ανιχνευθούν μη-τετριμμένες κοινότητες το null model πρέπει να περιέχει και το τυχαίο τμήμα αλλά και την «καθολική λειτουργία» του συστήματος.[23]

$$P_{ij} = C_{ij}^{(r)} + C_{ij}^{(m)}$$

2.5. Μέθοδος μεγιστοποίησης modularity.

Η συνάρτηση modularity Q η οποία αρχικά εισήχθη για τον ορισμό ενός κριτήριου για την ολοκλήρωση του αλγορίθμου των Girvan και Newman γρήγορα εξελίχθηκε σε ένα βασικό στοιχείο πολλών μεθόδων εύρεσης κοινοτήτων. [2]

Υποθέτοντας ότι υψηλές τιμές modularity υποδεικνύουν καλές διαμερίσεις του δικτύου σε κοινότητες γίνεται αντιληπτό ότι η διαμέριση που αντιστοιχεί στην μέγιστη τιμή σε ένα συγκεκριμένο δίκτυο θα πρέπει να είναι η καλύτερη. Αυτό είναι το κύριο κίνητρο για την μεγιστοποίηση της συνάρτησης modularity, της πιο γνωστής κλάσης μεθόδων για την εύρεση κοινοτήτων σε δίκτυα. Μία εξαντλητική βελτιστοποίηση της συνάρτησης modularity είναι αδύνατη λόγω του τεράστιου αριθμού τρόπων με τους οποίους μπορεί ένα δίκτυο να χωριστεί σε κοινότητες ακόμα και όταν αυτό είναι μικρό.[2]

Επιπλέον το πραγματικό μέγιστο της συνάρτησης δεν μπορεί να επιτευχθεί αφού πρόσφατα έχει αποδειχθεί ότι η βελτιστοποίηση της συνάρτηση modularity είναι ένα NP-complete πρόβλημα, δηλαδή είναι αδύνατον να βρεθεί η λύση σε χρόνο που αυξάνεται πολυωνυμικά με το μέγεθος του δικτύου. [2][24] Ωστόσο υπάρχουν αρκετοί αλγόριθμοι που μπορούν να βρουν μία ικανοποιητική προσέγγιση του μέγιστου της συνάρτησης σε εύλογο χρόνο.

2.5.1. Μέθοδος Louvain.

Η μέθοδος Louvain των Blondel, Guillaume, Lambiotte και Lefebvre [25] είναι ένας greedy αλγόριθμος, δηλαδή ακολουθεί ως λογική για την λύση του προβλήματος την επιλογή της τοπικά βέλτιστης επιλογής σε κάθε βήμα με σκοπό την εύρεση του

ολικού βέλτιστου. Σε πολλά προβλήματα μία greedy στρατηγική δεν παράγει την βέλτιστη λύση αλλά δίνει τοπικά βέλτιστη λύση η οποία προσεγγίζει την ολικά βέλτιστη λύση σε εύλογο χρόνο. [26]

Η μέθοδος Louvain είναι κατάλληλη για σταθμισμένα αλλά και για μη-σταθμισμένα δίκτυα. Ο αλγόριθμος χωρίζεται σε δύο βήματα που επαναλαμβάνονται συνεχώς. Έστω ένα σταθμισμένο δίκτυο με N κόμβους.

Αρχικά κάθε κόμβος του δικτύου αντιστοιχίζεται σε μία ξεχωριστή κοινότητα. Στη συνέχεια για κάθε κόμβο i ελέγχονται οι j γείτονές του και υπολογίζεται το κέρδος για τη συνάρτηση modularity στην περίπτωση που ο κόμβος i αφαιρεθεί από την κοινότητά του και τοποθετηθεί στην κοινότητα του κόμβου j. Ο κόμβος i τοποθετείται στην κοινότητα για την οποία το κέρδος είναι το μέγιστο, αλλά μόνο αν αυτό το κέρδος είναι θετικό. Εάν δεν υπάρχει θετικό κέρδος τότε ο κόμβος i παραμένει στην κοινότητα όπου ήταν αρχικά. Αυτή η διαδικασία επαναλαμβάνεται με τη σειρά για όλους τους κόμβους μέχρι να μην μπορεί να επιτευχθεί περαιτέρω βελτίωση. Το πρώτο βήμα σταματάει όταν ένα τοπικό μέγιστο επιτευχθεί για το modularity, δηλαδή καμία άλλη αλλαγή κόμβου δεν μπορεί να βελτιώσει την τιμή του modularity.

Πρέπει να σημειωθεί ότι το αποτέλεσμα της μεθόδου εξαρτάται από τη σειρά με την οποία θα ελεγχθούν οι κόμβοι. Έλεγχοι σε ήδη μελετημένα δίκτυα έχουν δείξει ότι η σειρά με την οποία θα ελεγχθούν οι κόμβοι δεν έχει σημαντική επίδραση στην μέγιστη τιμή modularity που δίνει ως αποτέλεσμα η μέθοδος, ωστόσο μπορεί να επηρεάσει τον χρόνο υπολογισμού.

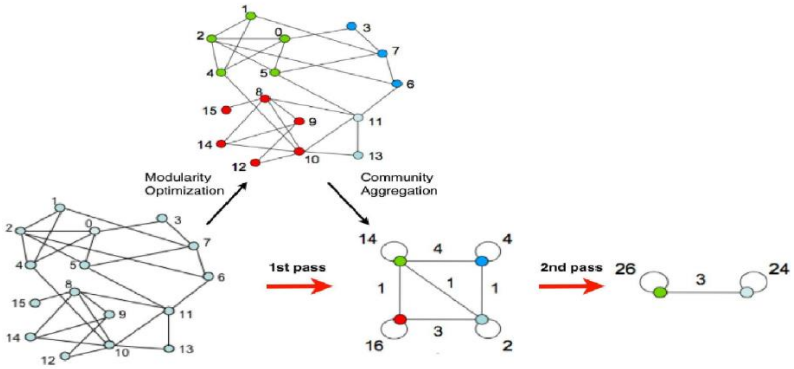
Η αποτελεσματικότητα του αλγορίθμου οφείλεται στο γεγονός ότι το κέρδος στην συνάρτηση modularity ΔQ από την μεταφορά ενός κόμβου από την κοινότητά του σε μία κοινότητα C μπορεί να υπολογιστεί εύκολα ως εξής:

$$\Delta Q = \left[\frac{\sum_{in}+k_{i,in}}{2m} - \left(\frac{\sum_{tot}+k_i}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right],$$

όπου $\sum_{in}$ είναι το άθροισμα των βαρών των εσωτερικών συνδέσεων των κόμβων της κοινότητας C, $\sum_{tot}$ είναι το άθροισμα των βαρών όλων των συνδέσεων που προσπίπτουν στους κόμβους της κοινότητας C, k_i είναι το άθροισμα των βαρών όλων των συνδέσεων του κόμβου i, $k_{i,in}$ είναι το άθροισμα των βαρών των συνδέσεων που ενώνουν τον κόμβο i με κόμβους της κοινότητας C, και m είναι το άθροισμα των βαρών όλων των συνδέσεων του δικτύου. Μία παρόμοια έκφραση χρησιμοποιείται για να υπολογιστεί και η αλλαγή στη τιμή modularity όταν ο κόμβος i αφαιρείται από την κοινότητά του. Στην πράξη υπολογίζεται η διαφορά στην τιμή modularity όταν ο

κόμβος i αφαιρείται από την κοινότητά του και στη συνέχεια προστίθεται σε μία γειτονική κοινότητα.

Το δεύτερο βήμα της μεθόδου αποτελείται από τη δημιουργία ενός καινούργιου δικτύου οι κόμβοι του οποίου αντιπροσωπεύουν τις κοινότητες που βρέθηκαν στο πρώτο βήμα. Τα βάρη των συνδέσεων μεταξύ των καινούργιων κόμβων δίνονται από το άθροισμα των βαρών των συνδέσεων μεταξύ των κόμβων των δύο αντίστοιχων κοινοτήτων. Συνδέσεις μεταξύ κόμβων της ίδια κοινότητας δημιουργούν βρόχους για αυτή την κοινότητα στο νέο δίκτυο.



Εικόνα 1: Οπτική περιγραφή των βημάτων του αλγορίθμου Louvain. [25]

Όταν ολοκληρωθεί το δεύτερο βήμα της μεθόδου μπορεί να επαναληφθεί το πρώτο βήμα στο καινούργιο σταθμισμένο δίκτυο που παράχθηκε. Ορίζεται ως επανάληψη ο συνδυασμός αυτών των δύο βημάτων. Ο αριθμός των μετα-κοινοτήτων μειώνεται σε κάθε επανάληψη επομένως ο περισσότερος χρόνος υπολογισμού χρησιμοποιείται για το πρώτο πέρασμα. Οι επαναλήψεις συνεχίζουν μέχρις ότου δεν υπάρχει αλλαγή στη τιμή του modularity και επιτυγχάνεται η μέγιστη τιμή για το modulatiry.

Η μέθοδος περιλαμβάνει μία ιδέα ιεραρχίας αφού δημιουργούνται κοινότητες κοινοτήτων κατά τις επαναλήψεις. Το ύψος της ιεραρχικής δομής που κατασκευάζεται ορίζεται από τον αριθμό των επαναλήψεων.

Η μέθοδος Louvain έχει 3 σημαντικά πλεονεκτήματα. Πρώτον τα βήματα είναι ευκολονόητα και το αποτέλεσμα προκύπτει χωρίς επιτήρηση. Ο αλγόριθμος είναι αρκετά γρήγορος, η πολυπλοκότητά του είναι γραμμική σε τυπικά και αραιά δεδομένα. Αυτό οφείλεται στο γεγονός ότι το κέρδος για το modularity υπολογίζεται εύκολα και ότι ο αριθμός των κοινοτήτων μειώνεται δραστικά σε κάθε επανάληψη. Τρίτον το

πρόβλημα του ορίου ανάλυσης του modularity το οποίο θα αναλυθεί σε επόμενη ενότητα φαίνεται να παρακάμπτεται χάρη στην εγγενή πολυεπίπεδη φύση του αλγορίθμου.

Η μέθοδος Louvain έχει δοκιμαστεί σε ορόσημα δίκτυα (benchmark networks), δηλαδή τεχνητά δίκτυα τα οποία έχουν εκ των προτέρων γνωστή δομή κοινοτήτων και χρησιμοποιούνται για την σύγκριση των μεθόδων εύρεσης κοινοτήτων. Έχει την καλύτερη απόδοση μεταξύ των μεθόδων που βασίζονται στην μεγιστοποίηση της τιμής modularity τόσο σε απλής μορφής δίκτυα οδηγούς (benchmark) όπως το Newman-Girvan benchmark δίκτυο όσο και σε πιο σύνθετα όπως το LFR benchmark δίκτυο.[27]

2.6. Όριο ανάλυσης και παράμετρος ανάλυσης modularity.

2.6.1. Όριο ανάλυσης modularity.

Το πρόβλημα του ορίου ανάλυσης της συνάρτησης modularity αναλύθηκε αρχικά από τους Fortunato και Barthelemy. [28] Είναι ένα μεθοδολογικό ζήτημα που αντιμετωπίζει η μεγιστοποίηση της συνάρτησης modularity. Κάτω από συγκεκριμένες προϋποθέσεις μπορεί να είναι αδύνατο να ανιχνευθούν κοινότητες μικρότερες από κάποια κλίμακα ακόμα και αν αυτές είναι καλά ορισμένες, δηλαδή είναι πλήρη υποδίκτυα. [29]

Το null model του modularity υποθέτει ότι οποιοσδήποτε κόμβος i μπορεί να συνδεθεί με οποιονδήποτε κόμβο j και το αναμενόμενο πλήθος συνδέσεων μεταξύ τους είναι $P_{ij} = \frac{k_i k_j}{2m}$. Όμοια το αναμενόμενο πλήθος συνδέσεων μεταξύ δύο κοινοτήτων A και B με συνολικούς βαθμούς K_A και K_B είναι $P_{AB} = \frac{K_A K_B}{2m}$. Η διαφορά στη τιμή του modularity μεταξύ της διαμέρισης στην οποία οι κοινότητες A και B θεωρούνται ως μία και της διαμέρισης που θεωρούνται ξεχωριστές κοινότητες είναι $\Delta Q_{AB} = \frac{l_{AB}}{m} - \frac{K_A K_B}{2m^2}$, όπου l_{AB} ο αριθμός των συνδέσεων που ενώνουν τοι κοινότητες A και B. Εάν $l_{AB} = 1$ δηλαδή υπάρχει μία μόνο σύνδεση μεταξύ των κόμβων της κοινότητας A και της κοινότητας B είναι αναμενόμενο ότι οι δύο κοινότητες δεν θα ενωθούν. Αντίθετα εάν $\frac{K_A K_B}{2m} < 1$, τότε $\Delta Q_{AB} > 0$. Έστω χάριν απλότητας ότι $K_A \sim K_B = K$, δηλαδή τα δύο υποδίκτυα έχουν περίπου το ίδιο πλήθος ακμών. Επομένως εάν $K <\sim \sqrt{2m}$, όπου m το συνολικό πλήθος των συνδέσεων του δικτύου και τα δύο υποδίκτυα συνδέονται η τιμή του modularity μεγαλώνει αν θεωρηθούν στην ίδια κοινότητα. [2][28]

Ο λόγος είναι προφανής, εάν υπάρχουν περισσότερες συνδέσεις από ότι αναμένονται μεταξύ των κοινοτήτων A και B, υπάρχει ισχυρή τοπολογική συσχέτιση μεταξύ των δύο κοινοτήτων. Εάν τα δύο υποδίκτυα είναι επαρκώς μικρά σε βαθμό, τότε ο αναμενόμενος αριθμός συνδέσεων από το null model μπορεί να είναι μικρότερος από ένα και έτσι ακόμα και η μικρότερη δυνατή σύνδεση επαρκεί για να συγχωνευτούν σε μία κοινότητα. Αυτό το αποτέλεσμα ισχύει ανεξάρτητα από την δομή των κοινοτήτων A και B, ακόμη και στην περίπτωση όπου τα δύο υποδίκτυα είναι κλίκες δηλαδή έχουν τη μέγιστη δυνατή πυκνότητα εσωτερικών συνδέσεων. [2]

Επομένως η βελτιστοποίηση της συνάρτησης modularity έχει ένα όριο ανάλυσης το οποίο μπορεί να εμποδίσει την ανίχνευση κοινοτήτων που είναι συγκριτικά μικρότερες από το δίκτυο σαν ολότητα, ακόμα και όταν είναι καλά ορισμένες. Επομένως εάν η διαμέριση του δικτύου που αντιστοιχεί στην μέγιστη τιμή modularity περιέχει κοινότητες με συνολικό βαθμό της τάξης μεγέθους $\sqrt{m}$ δεν γίνεται να γνωρίζουμε εκ των προτέρων αν οι κοινότητες είναι συνδυασμός μικρότερων και ασθενώς συνδεδεμένων μεταξύ τους κοινοτήτων ή όχι. [2]

Το όριο ανάλυσης του modularity πηγάζει από τον ορισμό του null model. Το μειονέκτημά του είναι ότι θεωρεί κάθε κόμβο ικανό να αλληλοεπιδράσει με οποιοδήποτε άλλο κόμβου του δικτύου, δηλαδή υπαινίσσεται ότι κάθε μέρος του δικτύου γνωρίζει για όλα τα υπόλοιπα. Αυτό ωστόσο είναι υπό αμφισβήτηση, και όπως έχει αποδειχθεί δεν ισχύει για μεγάλα συστήματα. Η υπόθεση ότι κάθε κόμβος έχει έναν περιορισμένο ορίζοντα μέσα στο δίκτυο και αλληλοεπιδρά με ένα μέρος αυτού είναι πιο λογική. [2]

2.6.2. Παράμετρος ανάλυσης modularity.

Για να παρακαμφθεί το πρόβλημα του ορίου ανάλυσης έχουν προταθεί διάφορες τεχνικές πολλαπλών ορίων (multiresolution). [30] Οι τεχνικές αυτές ενσωματώνουν στην ανάλυσή τους μία παράμετρο ανάλυσης στη έκφραση του modularity η οποία μπορεί να ρυθμιστεί για την ανίχνευση κοινοτήτων διαφορετικών μεγεθών. Στην φόρμουλα των Reichardt και Bornholdt για το modularity η παράμετρος ανάλυσης γ ορίζει τη σημαντικότητα του null model:

$$Q(\gamma) = \frac{1}{2m}\sum_{ij}(A_{ij} - \gamma P_{ij})\delta(C_i, C_j)$$

Όταν $\gamma < 1$ έχει ως αποτέλεσμα την εύρεση μεγαλύτερων κοινοτήτων ενώ όταν $\gamma > 1$ περισσότερες κοινότητες μικρότερου μεγέθους. Είναι σημαντικό να σημειωθεί

ότι μεταβάλλοντας την παράμετρο ανάλυσης γ επιτυγχάνεται η ανίχνευση κοινοτήτων διαφορετικών μεγεθών, αλλά δεν επιλύεται το πρόβλημα του ορίου ανάλυσης. Για κάθε τιμή της παραμέτρου γ το όριο ανάλυσης υπάρχει καθιστώντας αδύνατη την ανίχνευση κοινοτήτων συγκεκριμένου μεγέθους. Ένα πρόβλημα είναι ότι συνήθως δεν υπάρχει εκ των προτέρων πληροφορία για το μέγεθος των κοινοτήτων επομένως δεν είναι δυνατή η επιλογή της κατάλληλης τιμής της παραμέτρου γ. [2][29]

2.7. Στατιστική σημαντικότητα δομής κοινοτήτων.

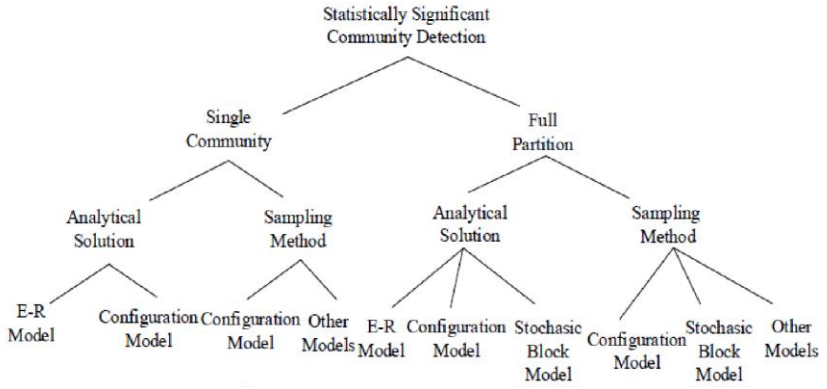
Η σημαντικότητα του προβλήματος εύρεσης κοινοτήτων σε δίκτυα έχει οδηγήσει στη δημιουργία πολλών αλγορίθμων οι οποίοι μπορούν να κατηγοριοποιηθούν ανάλογα με την ποιοτική συνάρτηση που χρησιμοποιούν για την αξιολόγηση της διαμέρισης που προκύπτει και την αντίστοιχη διαδικασία που ακολουθείται. Ωστόσο οι ποιοτικές συναρτήσεις δεν απαντούν στο ερώτημα της στατιστικής σημαντικότητας των κοινοτήτων. Δηλαδή στο πως θα αξιολογηθεί μία κοινότητα ή μία διαμέριση του δικτύου εάν είναι αληθής ή όχι, βάση ενός αυστηρού ελέγχου στατιστικής σημαντικότητας. Αριθμητικές τιμές ποιοτικών συναρτήσεων είναι συνήθως εξαρτώμενες από τα δεδομένα και επομένως η ερμηνεία και ορισμός ενός καθολικού κατωφλιού για όλα τα σύνολα δεδομένων είναι αδύνατη. Αντιθέτως η p-τιμή εξάγεται κάτω από ένα συγκεκριμένο μοντέλο τυχαίου δικτύου με μαθηματικά ορθό τρόπο. [31]

Για την ποσοτικοποίηση της στατιστικής σημαντικότητας της δομής κοινοτήτων ενός δικτύου πρέπει αρχικά να επιλεγεί ένα μοντέλο τυχαίου δικτύου. Τα μοντέλα τυχαίου δικτύου που έχουν χρησιμοποιηθεί περισσότερο στο πεδίο της στατιστικής σημαντικότητας κοινοτήτων είναι:

- Erdos-Renyi (E-R) μοντέλο, έχει δύο εκδοχές: $G(N, M)$ όπου ένα δίκτυο επιλέγεται τυχαία από το σύνολο όλων των δικτύων με N κόμβους και M συνδέσεις, $G(N, p)$ ένα δίκτυο N κόμβων κατασκευάζεται συνδέοντας δύο κόμβους τυχαία και ανεξάρτητα με πιθανότητα p, $p \in (0,1)$
- Configuration μοντέλο: δημιουργείται ένα τυχαίο δίκτυο στο οποίο κάθε κόμβος έχει σταθερό βαθμό, δηλαδή ο βαθμός ενός κόμβου είναι ίδιος σε οποιοδήποτε τυχαίο δίκτυο δημιουργηθεί.
- Στοχαστικό μπλοκ μοντέλο: παράγει ένα τυχαίο δίκτυο στο οποίο οποιοιδήποτε δύο κόμβοι συνδέονται με πιθανότητα η οποία ορίζεται από την κοινότητα στην οποία ανήκει ο κάθε ένας. Η πληροφορία για τις υποβόσκουσες κοινότητες

θεωρείται γνωστή και δύο κόμβοι θα συνδεθούν με μεγαλύτερη πιθανότητα εφόσον ανήκουν στην ίδια κοινότητα.

Δύο ακόμα κριτήρια εκτός από την επιλογή του μοντέλου τυχαίου δικτύου είναι ο στόχος της αξιολόγησης και η τεχνική για την εξαγωγή της p-τιμής. Ο στόχος της αξιολόγησης μπορεί να είναι είτε ολόκληρη η διαμέριση του δικτύου σε κοινότητες είτε μία συγκεκριμένη κοινότητα. Η διαφορά στην τεχνική της εξαγωγής της p-τιμής έγκειται στο αν θα χρησιμοποιηθούν δειγματοληπτικές τεχνικές ή αναλυτικές μέθοδοι. Το αντικείμενο της παρούσας εργασίας περιορίζεται στις δειγματοληπτικές μεθόδους οι οποίες θα αναλυθούν παρακάτω.



Εικόνα 2: Διαχωρισμός των υφιστάμενων μεθόδων για την εκτίμηση της p-τιμής.[31]

2.7.1. Στατιστική σημαντικότητα διαμέρισης.

2.7.1.1. z-score modularity.

Μία πρώτη προσέγγιση για τον υπολογισμό της στατιστικής σημαντικότητας μίας διαμέρισης του δικτύου σε κοινότητες είναι η σύγκριση της ποιότητας της δομής κοινοτήτων που εκτιμάται σε ένα πραγματικό δίκτυο με το ίδιο μέτρο σε ένα σύνολο τυχαίων δικτύων που διατηρούν την βαθμική ακολουθία του πραγματικού δικτύου αλλά οι συνδέσεις έχουν τοποθετηθεί τυχαία. [29]

Μεγιστοποιώντας το modularity σε ένα σύνολο τυχαίων δικτύων είναι δυνατόν να υπολογιστεί η μέση τιμή μ και η τυπική απόκλιση σ του modularity για αυτά και στη συνέχεια να συγκριθεί με τη τιμή modularity Q_0 της βέλτιστης διαμέρισης του πραγματικού δικτύου υπολογίζοντας το z score

$$z = \frac{Q_0 - \mu}{\sigma}$$

Το z-score μετράει σε μονάδες τυπικής απόκλισης πόσο απέχει η πραγματική τιμή από την μέση τιμή των τυχαίων δικτύων. Εάν $z \gg 1$ τότε η πραγματική τιμή modularity είναι σημαντικά μεγαλύτερη από τις τιμές των τυχαίων δικτύων.

Αυτή η προσέγγιση αντιμετωπίζει μερικά προβλήματα. Αρχικά δεν γνωρίζουμε εάν η εμπειρική κατανομή του modularity ακολουθεί κανονική κατανομή. Επίσης με αυτή τη μέθοδο παράγονται ψευδώς θετικά και ψευδώς αρνητικά αποτελέσματα (σφάλματα τύπου I και II). Υπάρχουν δίκτυα τα οποία δεν έχουν ισχυρή δομή κοινοτήτων και εμφανίζουν modularity σημαντικά μεγαλύτερο από τα τυχαία δίκτυα. Ακόμα υπάρχουν δίκτυα που παρότι έχουν ισχυρή δομή κοινοτήτων, το modularity που παρουσιάζουν δεν είναι σημαντικά μεγαλύτερο από τα τυχαία δίκτυα. [32]

2.7.1.2. Ανάλυση ανθεκτικότητας δικτύου.

Μία διαφορετική προσέγγιση είναι ο έλεγχος ανθεκτικότητας (robustness) του δικτύου και εμπεριέχει την πρόσθεση θορύβου στη διαδικασία εύρεσης κοινοτήτων. Οι εκτιμήσεις για τη δομή κοινοτήτων για ένα δίκτυο με καλά ορισμένες κοινότητες θα είναι παρόμοιες πριν και μετά την διαταραχή του δικτύου, ενώ για ένα δίκτυο με εύθραυστη δομή κοινοτήτων οι αλλαγές θα είναι μεγαλύτερες. [29]

Ένας έλεγχος ανθεκτικότητας περιλαμβάνει τα ακόλουθα βήματα:

1. Εύρεση μίας διαμέρισης C ενός πραγματικού δικτύου με N κόμβους και m συνδέσεις με κάποιον αλγόριθμο εύρεσης κοινοτήτων M.
2. Έστω ένα επίπεδο διαταραχής $p \in [0,1]$, το δίκτυο διαταράσσεται ανακατεύοντας τις συνδέσεις του, κάθε σύνδεση αφαιρείται με πιθανότητα p και αντικαθίσταται με μία καινούργια σύνδεση μεταξύ των κόμβων (i,j) που επιλέγονται τυχαία με πιθανότητα e_{ij}/m, όπου $e_{ij} = \frac{k_i k_j}{2m}$.
3. Χρησιμοποιώντας την ίδια μέθοδο M προκύπτει η διαμέριση C' για το διαταραγμένο δίκτυο και υπολογίζεται η variation of information $VI(C, C')$.
4. Επαναλαμβάνονται τα βήματα 2-3 για διαφορετικά επίπεδα διαταραχής $p \in [0,1]$ για να αποκτηθεί η καμπύλη VI_C. ($p = 0$ αντιστοιχεί στο πραγματικό δίκτυο, $p = 1$ αντιστοιχεί στο μέγιστα διαταραγμένο δίκτυο)
5. Επαναλαμβάνονται τα βήματα 1-4 με τη διαφορά ότι ως αρχικό δίκτυο στο βήμα 1 χρησιμοποιείται ένα δίκτυο το οποίο δημιουργήθηκε από κατάλληλο μοντέλο τυχαίου δικτύου (configuration model) και υπολογίζεται η καμπύλη $VI_{C_{random}}$

Η αξιολόγηση της σημαντικότητας της δομής κοινοτήτων του δικτύου C προκύπτει από τη σύγκριση των καμπυλών VI_C και $VI_{C_{random}}$. [32]

Στην βιβλιογραφία συναντώνται παραλλαγές της μεθόδου είτε χρησιμοποιώντας την αμοιβαία πληροφορία για την εκτίμηση της απόστασης των διαμερίσεων στο βήμα 3, [33] είτε προσθέτοντας ένα 6ο βήμα για τον έλεγχο της στατιστικής σημαντικότητας της διαφοράς μεταξύ της καμπύλης VI_C και $VI_{C_{random}}$. [34]

2.7.2. Στατιστική σημαντικότητα κοινότητας.

Η αντιμετώπιση του προβλήματος της αξιολόγησης της στατιστικής σημαντικότητας μίας κοινότητας περιλαμβάνει την παραγωγή τυχαίων δικτύων και στη συνέχεια η p-τιμή για μία κοινότητα c ορίζεται ως η πιθανότητα εύρεσης «καλύτερων κοινοτήτων». Οι διαφορές στις μεθόδους της βιβλιογραφία έγκεινται στο πως ορίζονται οι «καλύτερες κοινότητες». [31]

Οι Spirin και Mirny [35] υπολογίζουν την p-τιμή για την κοινότητα c με n κόμβους και m συνδέσεις ως την πιθανότητα ύπαρξης περισσότερων από m συνδέσεις μεταξύ των ίδιων ακριβώς κόμβων στα τυχαία δίκτυα. Αν και η πιθανότητα αυτή μπορεί να είναι αρκετά μικρή ο αριθμός Ω_n όλων των πιθανών συγκρίσιμων κοινοτήτων με n κόμβους είναι τεράστιος. Για να ληφθεί αυτό υπόψιν υπολογίζεται η E-τιμή, $E = p\Omega_n$ ως ο αναμενόμενος αριθμός κοινοτήτων με n κόμβους και m (ή παραπάνω) συνδέσεις. Ο αριθμός των πιθανώς συγκρίσιμων κοινοτήτων εκτιμάται ως: $\Omega_n = \binom{N}{N(d>d_c)}$, όπου N είναι ο συνολικός αριθμός των κόμβων του δικτύου, $N(d > d_c)$ το πλήθος των κόμβων που έχουν βαθμό μεγαλύτερο από d_c, και d_c η διάμεσος των βαθμών των κόμβων της κοινότητας c. Με αυτό τον τρόπο η E-τιμή λαμβάνει υπόψιν της το πλήθος των κόμβων της κοινότητας c αλλά και των αριθμό των συνδέσεων του κάθε κόμβου.

Μία πιο σύνθετη προσέγγιση είναι ο ορισμός ενός στατιστικού ελέγχου βασισμένου στην ποιοτική συνάρτηση και στο μέγεθος της κοινότητας. [36] Έστω μία κοινότητα c με τιμή modularity q_0 και μέγεθος s_0. Η κοινότητα c κρίνεται στατιστικά σημαντική εάν η τιμή q_0 είναι μεγαλύτερη από την αντίστοιχη τιμή κοινοτήτων ίδιου μεγέθους που ανιχνευθήκαν σε τυχαία δίκτυα. Υπολογίζεται η πιθανότητα $P(q > q_0|s_0)$, δηλαδή η πιθανότητα ότι μία κοινότητα μεγέθους s_0 ανιχνεύθηκε στα τυχαία δίκτυα και είχε τιμή modularity q μεγαλύτερη από q_0.

Κεφάλαιο 3: Μέθοδοι τυχαιοποίησης δικτύων.

3.1. Εισαγωγή.

Ένα τυχαιοποιημένο δίκτυο είναι ένα γράφημα που παράγεται από κάποια τυχαία διαδικασία. Στην ανάλυση δικτύων ένα βασικό ερώτημα είναι αν το παρατηρούμενο δίκτυο είναι τυχαίο. Στα εμπειρικά δίκτυα η ερμηνεία των αποτελεσμάτων αντιμετωπίζει δυσκολίες.

Μια στρατηγική για την αντιμετώπιση είναι η σύγκριση με το αναμενόμενο αποτέλεσμα από κάποιο κατάλληλο μοντέλο τυχαίου δικτύου. Το αναμενόμενο αποτέλεσμα μπορεί να εξαχθεί είτε αναλυτικά είτε από δείγματα που λαμβάνονται από προσομοιώσεις Monte Carlo. Συγκρίνοντας ένα παρατηρούμενο δίκτυο με ένα σύνολο τέτοιων γραφημάτων επιτρέπει σε κάποιον να ανιχνεύσει αποκλίσεις από την τυχαιότητα στις ιδιότητες του αρχικού δικτύου. Τυχαία δίκτυα με προκαθορισμένη βαθμική ακολουθία έχουν χρησιμοποιηθεί ευρέως ως μοντέλα για πολύπλοκα δίκτυα.[37]

Τα πραγματικά δίκτυα έχουν συχνά πολύ πλατιά κατανομή βαθμού, έχοντας σαν αποτέλεσμα ο βαθμός να είναι ένα σημαντικό χαρακτηριστικό του κόμβου το οποίο πρέπει να διατηρηθεί σε οποιαδήποτε τυχαιοποιημένη έκδοση του δικτύου. Οι πραγματικές εφαρμογές συχνά απαιτούν στα τυχαία δίκτυα να μην υπάρχουν βρόχοι και πολλαπλές ακμές μεταξύ οποιουδήποτε ζευγαριού κόμβων. [38]

Οι μέθοδοι τυχαιοποίησης δικτύων είναι τεχνικές που χρησιμοποιούνται για τον έλεγχο της σημαντικότητας και δημιουργούν τυχαία δίκτυα που διατηρούν κάποιες ιδιότητες του πραγματικού δικτύου όπως τον συνολικό βαθμό ή τη συνολική ισχύς αν οι συνδέσεις είναι σταθμισμένες, ή τον βαθμό κάθε κόμβου και αντίστοιχα την ισχύς κάθε κόμβου για σταθμισμένο δίκτυο.

Η μηδενική υπόθεση ελέγχει H_0 αν το υπό εξέταση δίκτυο είναι τυχαίο σύμφωνα με τον περιορισμό κάποιων ιδιοτήτων του αρχικού δικτύου. Δημιουργώντας Β τυχαιοποιημένα δίκτυα και εφόσον οι μεταβλητές που μελετάμε είναι ανεξάρτητες ακολουθεί η πραγματοποίηση του ελέγχου για την μηδενική υπόθεση H_0 ότι το αρχικό δίκτυο είναι τυχαίο, οι κόμβοι είναι ασυσχέτιστοι και δεν παρουσιάζει δομή κοινοτήτων.

Ως στατιστικό του ελέγχου θεωρούμε κάποιο χαρακτηριστικό του δικτύου q. Το στατιστικό q υπολογίζεται στο αρχικό δίκτυο, έστω q_0 η τιμή του σε αυτό και στα B

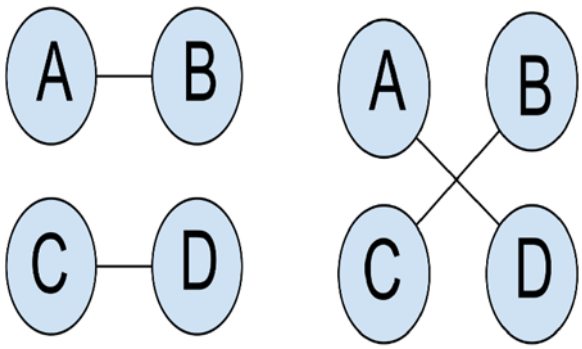
τυχαιοποιημένα δίκτυα με τιμές $q_1 \dots q_B$. Εξετάζουμε τη θέση r_0 της τιμής q_0 στο διατεταγμένο σύνολο $q_0, q_1, \dots, q_B$ και η p-τιμή του ελέγχου υπολογίζεται ως εξής:

$$\text{p-τιμή} = \begin{cases} \dfrac{2\mathrm{r}_0}{\mathrm{B}+1}, & \text{αν } \mathrm{r}_0 \leq \dfrac{\mathrm{B}+1}{2} \\ \dfrac{2(1-\mathrm{r}_0)}{\mathrm{B}+1}, & \text{αν } \mathrm{r}_0 \geq \dfrac{\mathrm{B}+1}{2} \end{cases}$$

3.2. Μέθοδος τυχαιοποίησης Maslov-Sneppen.

Ο αλγόριθμος των Maslov και Sneppen [39] είναι ο πιο δημοφιλής αλγόριθμος στη βιβλιογραφία τυχαιοποίησης δικτυών. Χρησιμοποιεί μία Μαρκοβιανή αλυσίδα για να παράξει ένα τυχαίο δίκτυο με δοσμένη βαθμική ακολουθία.

Η μέθοδος ξεκινάει από το αρχικό δίκτυο όπου ένα ζεύγος συνδέσεων $(A-B, C-D)$ επιλέγεται τυχαία και έχει ως αποτέλεσμα την αλλαγή των συνδέσεων ως εξής $(A-D, C-B)$ Αν οι καινούργιες συνδέσεις δεν υπάρχουν ήδη στο δίκτυο και δεν προκαλούν κάποιο βρόχο τότε γίνονται αποδέκτες και οι παλιές διαγράφονται. Εάν δεν γίνουν αποδεκτές τότε επιλέγεται τυχαία ένα άλλο ζεύγος συνδέσεων. Αυτή η διαδικασία επαναλαμβάνεται πολλές φορές για να προκύψει ένα τυχαιοποιημένο δίκτυο. Ενδεικτικά αναφέρεται ότι η προηγούμενή διαδικασία πρέπει να επαναληφθεί QE όπου E το πλήθος των συνδέσεων του αρχικού δικτύου και $Q = 100$. [38]



Εικόνα 3: : Βήμα της μεθόδου Maslov-Sneppen. [40]

Ο αλγόριθμος διατηρεί τον βαθμό κάθε κόμβου του αρχικού δικτύου στα τυχαιοποιημένα δίκτυα. Σε σταθμισμένα δίκτυα ο αλγόριθμος δεν διατηρεί την ισχύς κάθε κόμβου, αλλά την συνολική ισχύς του δικτύου. Σε κατευθυνόμενα δίκτυα ο αλγόριθμος διατηρεί είτε τον μέσα-βαθμό είτε τον έξω βαθμό των κόμβων.

Υπάρχουν δύο παραλλαγές του αλγόριθμου Maslov-Sneppen. Για να δημιουργηθεί η μέγιστη ιεραρχική έκδοση του αρχικού δικτύου πρέπει να προστεθεί μία συγκεκριμένη προτίμηση για την δημιουργία των καινούργιων συνδέσεων. Σε κάθε βήμα επιλέγεται δύο ζεύγη συνδεδεμένων κόμβων και προσπαθεί να συνδέσει τους δύο κόμβους με τους μεγαλύτερους βαθμούς. Οι υπόλοιποι 2 κόμβοι συνδέονται μεταξύ τους. Η μέγιστη αντι-ιεραρχική έκδοση του αρχικού δικτύου μπορεί να κατασκευαστεί με τον ίδιο αλγόριθμο αλλά με την αντίθετη προτίμηση για την δημιουργία των καινούργιων συνδέσεων, ο κόμβος με τον μεγαλύτερο βαθμό συνδέεται με τον κόμβο που έχει τον μικρότερο βαθμό. [41]

Ένα σημαντικό μειονέκτημα του αλγορίθμου είναι οι μεγάλες υπολογιστικές απαιτήσεις του. Ο αριθμός των βημάτων που απαιτούνται για να προκύψει ένα τυχαιοποιημένο δίκτυο είναι $O(L)$, όπου L είναι ο αριθμός των συνδέσεων και $O(L) = O(N)$ για αραιά δίκτυα ενώ $O(L) = O(N^2)$ για πυκνά δίκτυα. [42]

Εάν ο χρόνος που απαιτείται για να υπολογιστεί μία τοπική ιδιότητα του δικτύου είναι $O(N^t)$ τότε ο χρόνος που απαιτείται για να υπολογιστεί η ίδια ιδιότητα σε ένα σύνολο M τυχαιοποιημένων δίκτυών είναι $O(M \cdot L) + O(M \cdot N^t)$ δηλαδή $O(M \cdot N^t)$ εφόσον $t \geq 2$. [42]

3.3. Μέθοδος αλλαγής βαρών διατηρώντας την ισχύ.

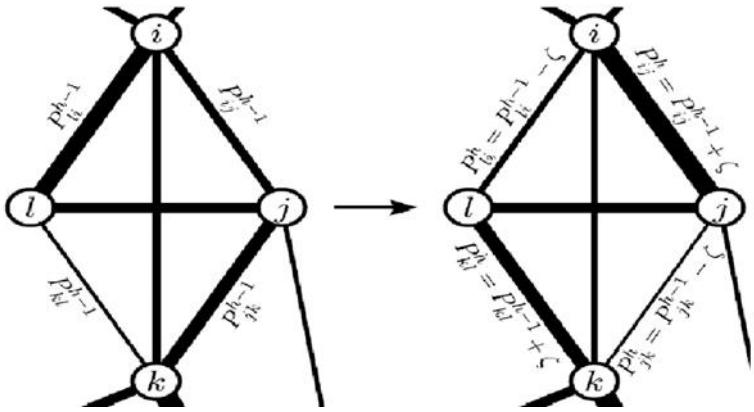
Η μέθοδος αυτή αναπτύχθηκε από τους Ansmann και Lehnertz και δοσμένου ενός αρχικού σταθμισμένου δικτύου παράγει τυχαιοποιημένα δίκτυα στα οποία η ισχύς κάθε κόμβου διατηρείται. [43]

Ένα μη-κατευθυνόμενο σταθμισμένο και πλήρες δίκτυο με n κόμβους μπορεί να εκφραστεί από ένα σύστημα n γραμμικών εξισώσεων με $\mathrm{m} := \frac{1}{2}\,\mathrm{n}(\mathrm{n}-1)$ βάρη συνδέσεων ως μεταβλητές. Δοσμένης της μη αρνητικότητας των βαρών των συνδέσεων το σύνολο λύσεων των εξισώσεων αντιπροσωπεύει ένα κυρτό πολύτοπο $\Omega \in \mathrm{R}^{\mathrm{m}}$του οποίου κάθε σημείο αντιστοιχεί σε ένα δίκτυο. Με αυτό τον τρόπο το πρόβλημα δημιουργίας ενός τυχαίου δικτύου που διατηρεί την ισχύς κάθε κόμβου είναι ισοδύναμο με την επιλογή τυχαίων σημείων από έναν πολύτοπο.

Η μέθοδος βασίζεται σε Hit-and-Run samplers, δηλαδή ένα σύνολο επαναληπτικών διαδικασιών Monte-Carlo που παράγουν δείγματα από μία φραγμένη περιοχή όπως ένας πολύτοπος. Η κατανομή των δειγμάτων ακολουθεί κατά προσέγγιση ομοιόμορφη κατανομή σε αυτή την περιοχή για μεγάλο αριθμό επαναλήψεων.

Τα βήματα της μεθόδου είναι τα εξής:

1. Επιλέγετε ένα δίκτυο $P^0 \in \Omega$ και ορίζεται ο μετρητής των επαναλήψεων $h = 1$.
2. Επιλέγονται τυχαία τέσσερα ζεύγη κόμβων με διακριτούς δείκτες $i, j, k, l \in \{1, \dots, n\}, (i, j), (i, l), (k, l), (j, k)$.
3. Επιλέγεται τυχαία ένας αριθμός ζ από την ομοιόμορφη κατανομή $[-min(P_{ij}^{h-1}, P_{kl}^{h-1}), min(P_{jk}^{h-1}, P_{li}^{h-1})]$. Έστω $P^h = P^{h-1}$ αλλά θέτουμε $P_{ij}^{h} = P_{ij}^{h-1} + \zeta$, $P_{kl}^{h} = P_{kl}^{h-1} + \zeta$, $P_{jk}^{h} = P_{jk}^{h-1} - \zeta$, $P_{li}^{h} = P_{li}^{h-1} - \zeta$
4. Αν $h < t$, αυξάνουμε το h κατά 1 και συνεχίζουμε στο βήμα 2. Σε αντίθετη περίπτωση το τυχαιοποιημένο δίκτυο είναι το P^h.



Εικόνα 4: Τετραγωνικός μετασχηματισμός του δικτύου Ph-1 στο δίκτυο Ph. [43]

Το διάστημα στο οποίο το ζ είναι περιορισμένο είναι το μέγιστο ένα, έτσι ώστε το τυχαιοποιημένο δίκτυο να μην περιέχει αρνητικά βάρη.

Το τυχαιοποιημένο δίκτυο P^t όπως παράγεται από τη διαδικασία αυτή είναι στατιστικά εξαρτημένο από το αρχικό δίκτυο P^0. Ωστόσο αυτή ή εξάρτηση γίνεται αμελητέα για κατάλληλα μεγάλο αριθμό επαναλήψεων t_{suf}. Σημαντική είναι η επιλογή του σημείου εκκίνησης P^0. Η πιο άμεση θα ήταν να διαλέξουμε ως P^0 το πραγματικό δίκτυο.

Ενδιαφέρον έχει και η επιλογή του κατάλληλου αριθμού επαναλήψεων. Στις περισσότερες περιπτώσεις η μέθοδος παράγει κατάλληλα τυχαιοποιημένα δίκτυα για $t = 2m$, όπου m είναι το πλήθος των συνδέσεων του αρχικού δικτύου.

Ο αλγόριθμος ωστόσο δεν είναι κατάλληλος για σταθμισμένα δίκτυα στα οποία τα βάρη είναι φραγμένα από κάποιο άνω όριο όπως τα δίκτυα συσχέτισης αφού ο συντελεστής συσχέτισης Pearson παίρνει τιμές στο διάστημα [-1,1].

Αυτό μπορεί να γίνει εύκολα κατανοητό μέσω ενός παραδείγματος. Για τιμές $P_{i,j}^{t} = 0{,}5$, $P_{j,k}^{t} = 0{,}3$, $P_{k,l}^{t} = 0{,}8$ και $P_{l,i}^{t} = 0{,}9$ ο αριθμός ζ θα επιλεγεί από την ομοιόμορφη κατανομή $U[-0{,}5\,,0{,}3]$. Για $\zeta = -0{,}5$ οι τιμές των συνδέσεων στην επόμενη επανάληψη $t+1$ θα είναι: $P_{i,j}^{t+1} = 0$, $P_{j,k}^{t+1} = 0{,}8$, $P_{k,l}^{t+1} = 0{,}3$ και $P_{l,i}^{t+1} = 1{,}4$.

3.4. Μέθοδος matching

Ο αλγόριθμος Matching είναι μία μέθοδος τυχαιοποίησης δικτύων που αναφέρεται σε κατευθυνόμενα δίκτυα. [38] Σε κάθε κόμβο εκχωρούνται ένα σύνολο από "stubs" δηλαδή κομμένα άκρα εισερχόμενων και εξερχόμενων συνδέσεων σύμφωνα με μία δοσμένη κατανομή έσω-βαθμού και έξω-βαθμού.

Στη συνέχεια οι έσω συνδέσεις και οι έξω συνδέσεις επιλέγονται τυχαία σε ζευγάρια και ενώνονται για να δημιουργήσουν τις τυχαιοποιημένες συνδέσεις του δικτύου. Αν μία πολλαπλή ακμή ή ένας βρόχος δημιουργηθεί τότε ολόκληρο το δίκτυο απορρίπτεται και η διαδικασία ξεκινάει από την αρχή. Η μέθοδος αυτή δημιουργεί επιτυχώς τυχαιοποιημένα κατευθυνόμενα δίκτυα με τις επιθυμητές ιδιότητες.

Ωστόσο πολλά πραγματικά δίκτυα έχουν κατανομές βαθμού με πλατιές ουρές που περιλαμβάνουν μία μειονότητα κόμβων με πολύ υψηλούς βαθμούς (hubs). Σε αυτή την περίπτωση το αναμενόμενο πλήθος ακμών μεταξύ δύο τέτοιων κόμβων ξεπερνάει τη μία καθιστώντας απίθανο η διαδικασία να ολοκληρωθεί και να δημιουργήσει κάποιο τυχαίο δίκτυο εκτός από αρκετά σπάνια περιπτώσεις.

Για να αντιμετωπιστεί αυτό το πρόβλημα μπορεί να χρησιμοποιηθεί μία τροποποίηση της μεθόδου όπου αν επιλεχθεί ένα ζεύγος που θα δημιουργήσει πολλαπλή ακμή ή βρόχο το δίκτυο να μην απορριφθεί αλλά να επιλεγεί ένα εναλλακτικό ζευγάρι τυχαία. Ο αλγόριθμος εισάγει μεροληψία και δημιουργεί ένα προκατειλημμένο δείγμα τυχαιοποιημένων δικτύων.

3.5. Μέθοδος go with the winners.

Ο αλγόριθμος αυτός είναι μία μη-Μαρκοβιανή αλυσίδα Monte Carlo για ομοιόμορφη δειγματοληψία από μία δοσμένη κατανομή. [38]

Ορίζεται μία οικογένεια M δικτύων. Η διαδικασία ξεκινάει εκχωρώντας τον κατάλληλο αριθμό "stubs" δηλαδή έσω συνδέσεων και έξω συνδέσεων για κάθε κόμβο. Επιλέγονται ανά ζεύγη τυχαία μία έσω σύνδεση με μία έξω σύνδεση και ενώνονται για να δημιουργήσουν μία ακμή.

Εάν δημιουργηθούν πολλαπλές ακμές ή βρόχοι το δίκτυο που τις περιλαμβάνει αφαιρείται από την οικογένεια M. Για να αντισταθμιστεί η μείωση του μεγέθους της οικογένειας, το μέγεθός της περιοδικά διπλασιάζεται με κλωνοποιώντας κάθε ένα από τα επιζώντα δίκτυα. Το στάδιο της κλωνοποίησης διεξαγάγετέ με προκαθορισμένο βαθμό έτσι ώστε να κρατήσει το μέγεθος της οικογένειας περίπου σταθερό κατά μέσο όρο.

Η διαδικασία αυτή επαναλαμβάνετε έως ότου να συνδεθούν όλες οι έσω συνδέσεις και έξω συνδέσεις, τότε ένα δίκτυο διαλέγεται τυχαία από την οικογένεια και του αναθέτεται ένα βάρος $W_i = -2^{-c}\frac{m}{M}$, όπου c είναι ο αριθμός των βημάτων κλωνοποίησης και m είναι ο αριθμός των επιζώντων δικτύων.

Η μέση τιμή οποιουδήποτε μεγέθους X σε ένα τέτοιο σύνολο δικτύων δίνεται από τη σχέση $\frac{\sum_i W_i X_i}{\sum_i W_i}$, όπου X_i η τιμή του X στο δίκτυο i. Η μέθοδος παράγει με ομοιόμορφο τρόπο γραφήματα από το δείγμα αλλά είναι σε σημαντικό βαθμό λιγότερο αποδοτική σε σχέση με τις μέθοδούς Maslov-Sneppen και Matching.

3.6. Μέθοδος μέγιστης πιθανοφάνειας.

Η μέθοδος βασίζεται στην εκτίμηση μέγιστης πιθανοφάνειας της μέγιστης εντροπίας μοντέλων δικτύων (Fagiolo, Squartini, & Garlaschelli, 2011). [42] Αντίθετα με άλλες αναλυτικές μεθόδους δεν απαιτεί παραδοχές για την δομή του αρχικού δικτύου όπως για παράδειγμα χαμηλή πυκνότητα.

Αρχικά καθορίζεται ένα επιθυμητό σύνολο τοπικών περιορισμών $\{C_a\}$. Υπολογίζεται αναλυτικά η πιθανότητα $P(G)$ υπό τους περιορισμούς $\{C_a\}$και μεγιστοποιεί την εντροπία $S \equiv -\sum_G P(G)\ln P(G)$ όπου G υποδηλώνει ένα συγκεκριμένο γράφημα στο δίκτυο και $P(G)$ την πιθανότητα να εμφανιστεί αυτό. Η πιθανότητα προσδιορίζει το σύνολο των δικτύων που χαρακτηρίζονται από τις επιθυμητές ιδιότητες. Τα δίκτυα G μπορεί να είναι δυαδικά ή σταθμισμένα και κατευθυνόμενα ή μη-κατευθυνόμενα.

Η επίσημη λύση της μεγιστοποίησης την εντροπίας μπορεί να γραφτεί από την Χαμιλτονιανή $H(G)$, η οποία αντιπροσωπεύει την ενέργεια ή το κόστος που συνδέεται με ένα δοσμένο γράφημα G. Η Χαμιλτονιανή ορίζεται ως ένας γραμμικός συνδυασμός των επιθυμητών περιορισμών $\{C_a\}$ και δίνεται από τον τύπο $\mathrm{H(G)} \equiv \sum_\alpha \theta_\alpha \mathrm{C}_\alpha(\mathrm{G})$, όπου $\{\theta_\alpha\}$ είναι ελεύθερες παράμετροι που λειτουργούν ως πολλαπλασιαστές Lagrange ελέγχοντας τις αναμενόμενες τιμές των περιορισμών σε όλο το σύνολο. Ο συμβολισμός $C_a(G)$ υποδηλώνει την συγκεκριμένη τιμή της ποσότητας C_a όταν αυτή μετριέται στο γράφημα G.

Από την Χαμιλτονιανή η μέγιστη εντροπία της πιθανότητας γραφήματος $P(G)$ μπορεί να δειχθεί ότι είναι $\mathrm{P(G)} = \frac{\mathrm{e}^{-\mathrm{H(G)}}}{\mathrm{Z}}$, όπου η ποσότητα Z για την κανονικοποίηση είναι η συνάρτηση κατανομής και ορίζεται ως $\mathrm{Z} = \sum_\mathrm{G} \mathrm{e}^{-\mathrm{H(G)}}$.

Στη συνέχεια μεγιστοποιείται η πιθανοφάνεια $P(G^*)$ ώστε να αποκτηθεί ένα συγκεκριμένο δίκτυο G^* το οποίο είναι το αρχικό δίκτυο προς τυχαιοποίηση. Αυτό το βήμα βελτιώνει τις τιμές των πολλαπλασιαστών Lagrange και μπορούμε να αποκτήσουμε αριθμητικές τιμές για τις μέσες αναμενόμενες τοπολογικές ιδιότητες στο σύνολο των τυχαιοποιημένων δικτύων.

Οι τιμές των παραμέτρων $\{\theta_\alpha\}$ που επιβάλουν τους τοπικούς περιορισμούς όπως αυτοί παρατηρούνται στο πραγματικό δίκτυο G^* βρίσκονται μεγιστοποιώντας το λογάριθμο της πιθανοφάνειας $\lambda \equiv \ln \mathrm{P(G^*)} = -\mathrm{H(G^*)} - \ln \mathrm{Z}$ για να αποκτήσουμε το πραγματικό δίκτυο G^*. Αυτή η σχέση είναι ισοδύναμη με την προϋπόθεση ότι το σύνολο των μέσων αναμενόμενων τιμών για κάθε περιορισμό C_a είναι ίσο με την εμπειρική τιμή που μετριέται στο πραγματικό δίκτυο.

Εφόσον οι τιμές των παραμέτρων έχουν προσδιοριστεί μπορεί να υπολογιστεί η αναμενόμενη τιμή οποιασδήποτε ιδιότητας X, $< \mathrm{X} > \equiv \sum_\mathrm{G} \mathrm{X(G)P(G)}$. Η ποσότητα $< X >$ αντιπροσωπεύει την μέση τιμή τις ιδιότητας X σε όλο το σύνολο των τυχαίων δικτύων με μέση τιμή των περιορισμών τους ίδια με το αρχικό δίκτυο.

Ενώ η μέθοδος Maslov-Sneppen παράγει ένα σύνολο δικτύων που περιέχει μόνο τα γραφήματα για τα οποία η τιμή κάθε περιορισμού C_a είναι ακριβώς ίση με την παρατηρούμενη τιμή $C_a(G^*)$ η μέθοδος μέγιστης πιθανοφάνειας παράγει ένα επεκταμένο σύνολο που περιλαμβάνει όλα τα πιθανά δίκτυα με N κόμβους και με μέση τιμή στο σύνολο για τον περιορισμό C_a ίση με την παρατηρούμενη τιμή στο πραγματικό δίκτυο $C_a(G^*)$. Μπορεί να δειχθεί ότι οι 2 μέθοδοι τείνουν προς σύγκλιση για μεγάλα δίκτυα.

Ο απαιτούμενος χρόνος για να αποκτηθεί η αναμενόμενη τιμή μιας ιδιότητας X σε όλο το σύνολο των τυχαιοποιημένων δικτύων είναι ο χρόνος που απαιτείται για να υπολογιστεί η ίδια ιδιότητα X στο πραγματικό δίκτυο δηλαδή $O(N^t)$ επομένως η μέθοδος μέγιστης πιθανοφάνειας είναι $O(M)$ φορές γρηγορότερη από την μέθοδο Maslov-Sneppen.

3.7. Μέθοδος τυχαιοποίησης με γραφική ακολουθία.

Ο αλγόριθμος τυχαιοποίησης βασίζεται σε μία άμεση χωρίς αλλαγές ακμών μέθοδο για τη συστηματοποιημένη δημιουργία όλων των απλών γραφημάτων δοσμένης μιας γραφικής ακολουθίας D. [44]

Ο αριθμός των στοιχείων του συνόλου $G(D)$ όλων των γραφημάτων που πραγματοποιούν μία γραφική ακολουθία D αυξάνει πολύ γρηγορά με το N. Επομένως για μεγάλες γραφικές ακολουθίες είναι απαραίτητο να γίνει δειγματοληψία του συνόλου $G(D)$.

Με δεδομένη μία γραφική ακολουθία ο αλγόριθμος έχει πάντα ως αποτέλεσμα ένα απλό γράφημα σε πολυωνυμικό χρόνο χωρίς να υποφέρει από απορρίψεις. Παρόλο που τα αποκτώμενα δείγματα δεν παράγονται ομοιόμορφα ο αλγόριθμος παρέχει για το ακριβές βάρος για κάθε δείγμα.

Έστω D μία μη-αύξουσα γραφική ακολουθία. Αρχικά επιλέγεται ο πρώτος κόμβος της ακολουθίας ως "hub" ή αλλιώς κεντρικός κόμβος και κατασκευάζεται το σύνολο των επιτρεπόμενων κόμβων $A = \{a_1, a_2, \dots, a_k\}$ που μπορούν να συνδεθούν με αυτόν έτσι ώστε να διατηρηθεί η γραφικότητα δηλαδή να μην υπάρχουν πολλαπλές ακμές ή βρόχοι.

Διαλέγεται τυχαία ένας κόμβος $\alpha \in A$ και δημιουργείται η σύνδεση μεταξύ του hub και του α. Εάν ο α έχει και άλλες συνδέσεις για να τοποθετηθούν τότε τον προσθέτουμε στο σύνολο X των απαγορευμένων κόμβων που περιέχει όλους τους κόμβους οι οποίοι δεν μπορούν να συνδεθούν πλέον με τον hub. Αλλιώς εάν ο α δεν έχει άλλες ακμές για να τοποθετηθούν αποκλείεται από περεταίρω μελέτη.

Η διαδικασία κατασκευής του A επαναλαμβάνεται και συνδέεται ο hub με τυχαία επιλεγμένους κόμβους από αυτό μέχρις ότου να μην έχει άλλες συνδέσεις. Τότε επιλέγεται ο επόμενος hub και επαναλαμβάνονται τα βήματα μέχρι να τοποθετηθούν όλες οι συνδέσεις και να προκύψει το τυχαίο δίκτυο.

Κάθε φορά που η διαδικασία επαναλαμβάνεται η βαθμική ακολουθία D που λαμβάνουμε υπόψη είναι η βαθμική ακολουθία των υπολοίπων δηλαδή η πραγματική βαθμική ακολουθία μειωμένη από τις συνδέσεις που έχουν πραγματοποιηθεί στο προηγούμενο βήμα και χωρίς να περιλαμβάνει κόμβους με μηδενικό βαθμό.

Το βάρος για το τυχαίο δίκτυο έτσι ώστε να αποκτήσουμε μη-μεροληπτικές εκτιμήσεις είναι οι αντεστραμμένες δεσμευμένες πιθανότητες παραγωγής του συγκεκριμένου δικτύου.

Στη διαδικασία κατασκευής του τυχαίου δικτύου m διαφορετικοί κόμβοι $i = 1,2,\dots,m$ έχουν επιλεγεί ως hub και έχουν d_i υπολειπόμενο βαθμό όταν επιλέχθηκαν. Έτσι το βάρος για ένα συγκεκριμένο τυχαίο δίκτυο δίνεται από τον τύπο $\mathrm{w} = \Pi_{\mathrm{i}=1}^{\mathrm{m}} \frac{1}{\mathrm{d_i!}} \Pi_{\mathrm{j}=1}^{\mathrm{d_i}} \mathrm{k_{i_j}}$, όπου $\mathrm{k_{i_j}}$ τα μεγέθη των συνόλων επιτρεπόμενων κόμβων A.

Το βάρος εξηγούνται αν λάβουμε υπόψιν το γεγονός ότι σε κάθε βήμα ο κόμβος hub έχει $\mathrm{k_{i_j}}$ επιλογές κόμβων με τους οποίους μπορεί να συνδεθεί, δηλαδή το μέγεθος του συνόλου A, και ότι ο αριθμός των ισοδύναμων τρόπων για να συνδεθεί ένας νέος hub είναι $d_i!$.

Κεφάλαιο 4: Μέθοδοι τυχαιοποίησης δικτύων συσχέτισης.

4.1. Εισαγωγή.

Τα δίκτυα έχουν χρησιμοποιηθεί στην ανάλυση πολυμεταβλητών χρονοσειρών. Έχουν ως κόμβους τις παρατηρούμενες μεταβλητές και οι συνδέσεις τους καθορίζονται από κάποιο μέτρο συσχέτισης για μη κατευθυνόμενες συνδέσεις για παράδειγμα συσχέτιση Pearson ή κάποιο μέτρο αιτιότητας για κατευθυνόμενες συνδέσεις όπως η αιτιότητα Granger.

Στην ανάλυση δικτύων ένα ενδιαφέρον ερώτημα σε πολλές περιπτώσεις είναι εάν το παρατηρούμενο δίκτυο ξεχωρίζει από ένα τυχαίο δίκτυο με την ίδια πυκνότητα συνδέσεων. Για να απαντηθεί αυτό το ερώτημα πρώτα δημιουργούνται τυχαία δίκτυα εναλλάσσοντας τις συνδέσεις και διατηρώντας τον βαθμό ή την ισχύς κάθε κόμβου. Το πραγματικό δίκτυο συγκρίνεται με αυτά τα δίκτυα με χρήση κατάλληλων δεικτών.

Η ανάλυση αυτή έχει αποδειχθεί ότι δεν είναι κατάλληλη για δίκτυα πολυμεταβλητών χρονοσειρών καθώς προσομοιώσεις σε πολυμεταβλητές χρονοσειρές έχουν δείξει ότι η κλασική τυχαιοποίηση δικτύων εσφαλμένα τείνει να απορρίπτει την μηδενική υπόθεση για το τυχαίο δίκτυο. [45]

Για τα δίκτυα που παράγονται από πολυμεταβλητές χρονοσειρές η μηδενική υπόθεση H_0 του τυχαίου δικτύου μπορεί να έχει ένα διαφορετικό νόημα. Η τυχαιότητα του δικτύου μπορεί να μην αναφέρεται στην τοπολογία του αλλά στην ανεξαρτησία των παρατηρούμενων μεταβλητών.

Ένα δίκτυο συσχέτισης παράγεται από έναν πίνακα συσχέτισης επομένως έχει κάποια δομή από τον τρόπο κατασκευής του ακόμα και όταν οι μεταβλητές είναι τελείως ασυσχέτιστες. Έπειτα ένα δίκτυο συσχέτισης που παράγεται από ανεξάρτητες μεταξύ τους χρονοσειρές δεν μπορεί να είναι τελείως τυχαίο διότι ο πίνακας συσχέτισης είναι πάντα θετικά ημι-ορισμένος δηλαδή οι ιδιοτιμές του είναι μη-αρνητικές. [45]

Μία άλλη ιδιότητα της μη-τυχαίας τοπολογίας των δικτύων συσχέτισης είναι μοτίβο μεταβατικότητας. Ισχυρή θετική συσχέτιση στο μη-άμεσο μονοπάτι ($X - Y - Z$) υπαινίσσεται συσχέτιση στην άμεση σύνδεση ($X - Z$) η οποία είναι σημαντικά μεγαλύτερη από την αναμενόμενη σε τυχαία δίκτυα (παρόλο που δεν σημαίνει ότι τα μη-άμεσα μονοπάτια προβλέπουν άμεσες συνδέσεις σε όλες τις περιπτώσεις. [45]

Συγκεκριμένα δίκτυα οι συνδέσεις των οποίων δημιουργούνται με κάποιο μέτρο συσχέτισης είναι εγγενώς πιο ομαδοποιημένα από τα τυχαία δίκτυα, ενώ αυτά που

δημιουργούνται χρησιμοποιώντας μέτρα μερικής συσχέτισης είναι λιγότερο ομαδοποιημένα από τα τυχαία δίκτυα. [46]

4.2. Υποκατάστατα δεδομένα και έλεγχος υπόθεσης υποκατάστατων δεδομένων.

Η μέθοδος δημιουργίας υποκατάστατων δεδομένων δημιουργήθηκε από τον Theiler και χρησιμοποιείται για την μη γραμμική ανάλυση χρονοσειρών. Τα υποκατάστατα δεδομένα είναι τυχαίες παραλλαγές των αρχικών δεδομένων κάτω από μία μηδενική υπόθεση H_0.

Ο έλεγχος υπόθεσης υποκατάστατων δεδομένων είναι ευρέως διαδεδομένος σε σχέση με την μηδενική υπόθεση ότι η υπό εξέταση χρονοσειρά παράγεται από μία Γκαουσιανή γραμμική διαδικασία που υποβάλετε σε έναν ενδεχομένως μη γραμμικό μετασχηματισμό. Κατάλληλα σχεδιασμένα υποκατάστατα δεδομένα για αυτή τη μηδενική υπόθεση πρέπει να έχουν την ίδια αυτοσυσχέτιση και το ίδιο ιστόγραμμα με τα πραγματικά δεδομένα.

Όταν δεν προσδιορίζεται μία εναλλακτική υπόθεση, όπως στην περίπτωση του ελέγχου για τη μη γραμμικότητα είναι δύσκολο να βρεθεί ένα στατιστικό για τον έλεγχο με γνωστή αναλυτική κατανομή. Επομένως χρησιμοποιούνται προσομοιώσεις Monte Carlo για να δημιουργηθεί η εμπειρική κατανομή του επιλεγμένου στατιστικού από τις αριθμητικές τιμές που παίρνει αυτό το στατιστικό για ένα σύνολο υποκατάστατων δεδομένων που αντιπροσωπεύουν την μηδενική υπόθεση. [47]

Ένας έλεγχος υποκατάστατων δεδομένων περιλαμβάνει τα βήματα που ακολουθούν:

1. Υπολογισμός ενός στατιστικού από τα πραγματικά δεδομένα με τιμή έστω Q_0.
2. Δημιουργία N υποκατάστατων συνόλων δεδομένων με βάση τα πραγματικά δεδομένα.
3. Υπολογισμός του ίδιου στατιστικού για κάθε σύνολο υποκατάστατων δεδομένων: $Q_1, Q_2, \dots, Q_N$.
4. Σχηματίζεται η ολοκληρωμένη λίστα με το πραγματικό και τα υποκατάστατα στατιστικά $\{Q_0, Q_1, Q_2, \dots, Q_N\}$, διατάσσεται με αύξοντα τρόπο με r_i να δηλώνει τη θέση του στοιχείου Q_i.

Ιδιαίτερο ενδιαφέρον έχει η θέση r_0 του στατιστικού Q_0 στην αύξουσα λίστα. Αν η μηδενική υπόθεση είναι αληθής, και η μέθοδος δημιουργίας των υποκατάστατων δεδομένων είναι καλή, τότε το στατιστικό Q_0 πρέπει να είναι δυσδιάκριτο από τα υποκατάστατα στατιστικά και η θέση του r_0 μπορεί να πάρει οποιαδήποτε τιμή από 1 έως $N + 1$.

Επομένως ένας μονόπλευρος έλεγχος που απορρίπτει την μηδενική υπόθεση όταν η θέση r_0 είναι μικρότερη ή ίση με κάποιο R_0 παράγει ψευδώς θετικά αποτελέσματα $a = \frac{R_0}{N+1}$. Η πιθανότητα παραγωγής ενός ψευδούς θετικού αποτελέσματος ονομάζεται μέγεθος του ελέγχου. Το «κατ' όνομα» μέγεθος του ελέγχου είναι η τιμή $\frac{R_0}{N+1}$, ενώ το «πραγματικό» μέγεθος του ελέγχου είναι το ποσοστό των ψευδώς θετικών αποτελεσμάτων που λαμβάνει κανείς στην πραγματικότητα.

Η ισχύς του ελέγχου είναι το ποσοστό των αληθώς θετικών αποτελεσμάτων, δηλαδή η πιθανότητα απόρριψης της μηδενικής υπόθεσης όταν η μηδενική υπόθεση δεν είναι αληθής. [48]

4.3. Δίκτυα συσχέτισης χρονοσειρών.

Στα δίκτυα συσχέτισης οι κόμβοι αντιστοιχούν σε τυχαίες μεταβλητές και οι συνδέσεις δίνονται από κάποιο στατιστικό μέτρο διασυσχέτισης όπως ο συντελεστής συσχέτισης Pearson.

Για ένα σύνολο K τυχαίων μεταβλητών $\{X_1, X_2, \dots, X_K\}$ δίνεται ένα δείγμα n χρονικά διατεταγμένων παρατηρήσεων για κάθε μεταβλητή $\{x_{k,1}, x_{k,2}, \dots, x_{k,n}\}$ για $k = 1, \dots, K$ οι οποίες μπορεί να είναι τελείως ανεξάρτητες, να υπάρχει εξάρτηση μεταξύ των παρατηρήσεων της ίδιας μεταβλητής (ύπαρξη σημαντικής αυτοσυσχέτισης) αλλά και μεταξύ των μεταβλητών (ύπαρξη σημαντικής διασυσχέτισης).

Ως μέτρο της αλληλεπίδρασης των μεταβλητών χρησιμοποιείται ο συντελεστής συσχέτισης Pearson χωρίς χρονική υστέρηση. Για δύο μεταβλητές $X = X_i$ και $Y = X_j$ $i, j \in \{1, \dots, K\}$ ο συντελεστής συσχέτισης Pearson δίνεται από τον τύπο $\mathrm{r_{X,Y}} = \frac{\mathrm{s_{X,Y}}}{\sqrt{\mathrm{s_X^2 s_Y^2}}}$,

όπου $\mathrm{s_{X,Y}} = \frac{1}{\mathrm{n-1}} \sum_{\mathrm{t=1}}^{\mathrm{n}} (\mathrm{x_t} - \bar{\mathrm{x}})(\mathrm{y_t} - \bar{\mathrm{y}})$ είναι η δειγματική συνδιασπορά των X και Y και $\mathrm{s_X^2} = \frac{1}{\mathrm{n-1}} \sum_{\mathrm{t=1}}^{\mathrm{n}} (\mathrm{x_t} - \bar{\mathrm{x}})^2$ η δειγματική διασπορά και $\bar{\mathrm{x}}$ η δειγματική μέση τιμή της μεταβλητής X.

Ο συντελεστής συσχέτισης Pearson παίρνει τιμές από -1 έως 1. Με αυτόν τον τρόπο δημιουργείται ο πίνακας συσχέτισης R και είναι συμμετρικός όταν αναφερόμαστε σε μη κατευθυνόμενο δίκτυο. Ο πίνακας συσχέτισης R είναι πάντα θετικά ημι-ορισμένος ανεξάρτητα από δομή συσχέτισης των τυχαίων μεταβλητών, ακόμα και όταν το δίκτυο είναι τελείως τυχαίο δηλαδή οι μεταβλητές είναι ασυσχέτιστες. [45]

Για τη δημιουργία μη-σταθμισμένου δικτύου ο πίνακας γειτνίασης A προκύπτει από τον πίνακα συσχέτισης R θέτοντας 1 όταν η συσχέτιση κριθεί στατιστικά σημαντική σύμφωνα με κάποιο κριτήριο και 0 αλλιώς. Το κριτήριο μπορεί να είναι ένα αυθαίρετο κατώφλι, ένα κατώφλι που να αντιστοιχίζεται σε μία συγκεκριμένη πυκνότητα συνδέσεων ή να είναι αποτέλεσμα ενός τεστ στατιστικής σημαντικότητας.[49]

Για τη δημιουργία σταθμισμένου δικτύου ο πίνακας βαρών W προκύπτει από κατάλληλη μετατροπή του πίνακα συσχέτιση R έτσι ώστε όλα τα στοιχεία του $w_{i,j}$ να είναι θετικά.

Υπάρχουν διαφορετικές προσεγγίσεις για τη δημιουργία του πίνακα βαρών W. Ένας τρόπος για την κατασκευή του W είναι να δημιουργηθεί από τις απόλυτες τιμές των στοιχείων του πίνακα συσχέτισης R, $\mathrm{w_{i,j}} = \left|\mathrm{r_{i,j}}\right|$. [49]

Η υποχρεωτική συνθήκη του θετικά ημι-ορισμένου πίνακα συσχέτισης R ορίζει τις επιτρεπτές μορφές που μπορούν να έχουν ο πίνακας γειτνίασης A και ο πίνακας βαρών W.

4.4. Μέθοδος δημιουργίας τυχαίου πίνακα συσχέτισης

Τυχαίοι πίνακες συσχέτισης μπορούν να δημιουργηθούν είτε με δειγματοληψία από την κατανομή Wishart είτε χρησιμοποιώντας τον αλγόριθμο Hirschberger-Qi-Steuer (H-Q-S). [50] Ο αλγόριθμος H-Q-S συνήθως έχει ως αποτέλεσμα μια κατανομή τιμών συσχέτισης η οποία προσεγγίζει καλύτερα με την πραγματική κατανομή τιμών συσχέτισης ενός πραγματικού δικτύου σε σύγκριση με τη δειγματοληψία από την κατανομή Wishart. Η επιτυχής προσέγγιση της πραγματικής κατανομής των τιμών συσχέτισης είναι σημαντική για να διασφαλιστεί η διατήρηση της πυκνότητας στο τυχαίο δίκτυο. [51]

Ο αλγόριθμος H-Q-S δημιουργεί τυχαίους πίνακες συνδιασπορών οι οποίοι ταιριάζουν με τις ιδιότητες της κατανομής ενός παρατηρούμενου πίνακα

συνδιασπορών. Βασίζεται στην διάσπαση Cholesky, η οποία αξιώνει ότι για κάθε πίνακα $X \in R^{N \times m}$, ο πίνακας XX^T είναι θετικά ημι-ορισμένος και επομένως μπορεί να λειτουργήσει ως πίνακας συνδιασπορών. Η βασική ιδέα που ακολουθεί είναι η ανεξάρτητη δειγματοληψία στοιχείων για τον πίνακα X από την Γκαουσιανή κατανομή. Η μέση τιμή μ και η διασπορά σ^2 της Γκαουσιανής κατανομής καθώς και η τιμή της παραμέτρου m επιλέγονται ώστε να διασφαλιστεί ότι οι ιδιότητες της κατανομής του παρατηρούμενου πίνακα συνδιασπορών διατηρούνται στο τυχαίο πίνακα συνδιασπορών XX^T. [46] Ο αλγόριθμος διατηρεί την μέση τιμή e, τη διασπορά v, των μη-διαγώνιων στοιχείων (συνδιασπορές) καθώς και την μέση τιμή $\hat{e}$ των διαγώνιων στοιχείων (διασπορές) του εμπειρικού πίνακα συνδιασπορών. [51] Ο αλγόριθμος H-Q-S χρησιμοποιεί ως είσοδο τις τιμές e, v, $\hat{e}$ και N αποτελείται από τα ακόλουθα βήματα:

- $m \leftarrow max(2, \lfloor \hat{e}^2 - e^2/v \rfloor)$
- $\mu \leftarrow \sqrt{e/v}$
- $\sigma^2 \leftarrow -\mu^2 + \sqrt{\mu^4 + v/m}$
- $x_{i,j} \sim N(\mu, \sigma^2), i = 1, \dots, N, j = 1, \dots, N$
- $C = XX^T$

Το γεγονός ότι ο αλγόριθμος αναφέρεται σε πίνακα συνδιασπορών δεν αποτελεί περιορισμό διότι μπορεί να μετασχηματιστεί σε πίνακα συσχέτισης με χρήση των τυπικών αποκλίσεων για κανονικοποίηση. Ο πίνακας συνδιασπορών C μετατρέπεται σε πίνακα συσχέτισης ως εξής: aCa όπου $a = diag\left(\frac{1}{\sqrt{C_{1,1}}}, \dots, \frac{1}{\sqrt{C_{N,N}}}\right)$. [46]

4.5. Μέθοδοι τυχαιοποίησης δικτύου μέσω τυχαιοποίησης πολυμεταβλητής χρονοσειράς.

Πέρα από μεθόδους τυχαιοποίησης δικτύων που διατηρούν την μεταβατική δομή ενός πίνακα συσχέτισης [46] μία διαφορετική προσέγγιση για την τυχαιοποίηση του πίνακα συσχέτισης προτείνει την τυχαιοποίηση της πολυμεταβλητής χρονοσειράς αντί του πίνακα συσχέτισης.

Αυτή η προσέγγιση έχει δεχθεί λίγη προσοχή στη βιβλιογραφία και οι πρώιμες εφαρμογές έχουν χρησιμοποιήσει είτε απλό ανακάτεμα των χρονοσειρών [52] ή τυχαιοποίηση φάσης. [46] [51]

Πιο πρόσφατη είναι η μέθοδος που έχει δημιουργηθεί από τους Χορόζογλου και Κουγιουμτζή [45] [53] όπου η τυχαιοποίηση ενός δικτύου συσχέτισης γίνεται

χρησιμοποιώντας μία κατάλληλη μέθοδο για την τυχαιοποίηση της πολυμεταβλητής χρονοσειράς. Οι τυχαιοποιημένες (υποκατάστατες) χρονοσειρές διατηρούν την δομή αυτοσυσχέτισης και είναι ασυσχέτιστες μεταξύ τους, έτσι ώστε το δίκτυο συσχέτισης που δημιουργείται από τις υποκατάστατες χρονοσειρές κληρονομεί όποιες ψευδείς συσχετίσεις υπήρχαν στο αρχικό δίκτυο συσχέτισης λόγω της ύπαρξης ισχυρής αυτοσυσχέτισης.

4.5.1. Μέθοδος δημιουργίας σταθμισμένου τυχαιοποιημένου δικτύου.

Η μέθοδος λαμβάνει υπόψιν της τους περιορισμούς για έναν πίνακα συσχέτισης R και επομένως για τον πίνακα βαρών W και τον πίνακα γειτνίασης A. Η λογική της μεθόδου είναι να τυχαιοποιηθούν οι χρονοσειρές αντί απευθείας των συνδέσεων του δικτύου συσχέτισης.

Κάθε μία από τις K χρονοσειρές $\{x_{k,t}\}_{t=1}^{n}, k = 1, \dots, K$ τυχαιοποιείται ξεχωριστά υπό την προϋπόθεση να διατηρηθεί η περιθώρια κατανομή και η συνάρτηση αυτοσυσχέτισης [54] ή ισοδύναμα το φάσμα ισχύος. [55]

Για την τυχαιοποίηση των χρονοσειρών χρησιμοποιείται ο αλγόριθμος Iterative Amplitutde Adjusted Fourrier Transform (IAAFT) των Schreiber kai Schmitz [55] διότι είναι κατάλληλος για πολύ μικρές χρονοσειρές και παρέχει την ελάχιστη διακύμανση για την προσέγγιση της αυτοσυσχέτισης. Επίσης χρησιμοποιείται ο αλγόριθμος STAP [54] ο οποίος έχει μικρότερη μεροληψία σε σχέση με τον αλγόριθμο IAAFT αλλά και μεγαλύτερη διακύμανση στην προσέγγιση της αυτοσυσχέτισης. Στη θέση των δύο αυτών αλγορίθμων μπορεί να χρησιμοποιηθεί κάποιος άλλος αλγόριθμος παραγωγής υποκατάστατων χρονοσειρών ανάλογα με τη φύση των χρονοσειρών. [45]

Η διαδικασία RTSweight για την παραγωγή B σταθμισμένων τυχαιοποιημένων δικτύων συσχέτισης έχει ως εξής:

1. Για κάθε χρονοσειρά $\{x_{k,t}\}_{t=1}^{n}, k = 1, \dots, K$ δημιουργούμε μία υποκατάστατη χρονοσειρά $\{x_{k,t}^{*}\}_{t=1}^{n}, k = 1, \dots, K$ χρησιμοποιώντας τον αλγόριθμο IAAFT και έτσι παράγεται η υποκατάστατη πολυμεταβλητή χρονοσειρά $\{x_{1,t}^{*}, \dots, x_{K,t}^{*}\}_{t=1}^{n}$.
2. Υπολογίζεται ο πίνακας συσχέτισης R^{*} για τις $\{x_{1,t}^{*}, \dots, x_{K,t}^{*}\}_{t=1}^{n}$
3. Κατασκευάζεται ο κατάλληλος πίνακας βαρών W^{*} από τον R^{*}.
4. Τα βήματα 1-3 επαναλαμβάνονται B φορές για να παραχθούν B τυχαιοποιημένα δίκτυα συσχέτισης.

Οι Κ χρονοσειρές $\{x^*_{1,t}, \dots, x^*_{K,t}\}^n_{t=1}$ είναι από κατασκευής ανεξάρτητες μεταξύ τους και οι μη-διαγώνιες τιμές του πίνακα συσχέτισης R^* πρέπει να είναι μηδέν και στην πραγματικότητα στατιστικά μη-σημαντικές. Ωστόσο είναι γνωστό ότι η ισχυρή αυτοσυσχέτιση μπορεί να έχει ως αποτέλεσμα αυξημένη ψευδή διασυσχέτιση μεταξύ των χρονοσειρών και επομένως οι τιμές διασυσχέτισης του πίνακα R^* να βρεθούν στατιστικά σημαντικές χρησιμοποιώντας κάποιο κατώφλι ή ένα τεστ στατιστικής σημαντικότητας. Έτσι στατιστικά σημαντικές συνδέσεις στο πραγματικό δίκτυο μπορεί να αντιστοιχίζονται σε ψευδείς τιμές διασυσχέτισης. [45]

Σύμφωνα με τη μέθοδο αυτή μία τέτοια ψευδής σύνδεση στο πραγματικό δίκτυο μπορεί να διατηρηθεί στο τυχαιοποιημένο δίκτυο αφού το συγκεκριμένο ζευγάρι υποκατάστατων χρονοσειρών διατηρεί τις αρχικές αυτοσυσχετίσεις.

Η μέθοδος δεν διατηρεί την ισχύ κάθε κόμβου ούτε την συνολική ισχύ του αρχικού δικτύου.

4.5.2. Μέθοδος δημιουργίας μη-σταθμισμένου τυχαιοποιημένου δικτύου.

Για την δημιουργία μη-σταθμισμένων τυχαιοποιημένων δικτύων η διαδικασία είναι παρόμοια με την δημιουργία σταθμισμένων τυχαιοποιημένων δικτύων με αλλαγή στο βήμα 3.

Υπάρχουν δύο διαφορετικές προσεγγίσεις σε συνάρτηση με τη διατήρηση ή όχι του πλήθους των συνδέσεων του αρχικού δικτύου.

4.5.2.1. Μη-σταθμισμένα τυχαιοποιημένα δίκτυα χωρίς διατήρηση του πλήθους των συνδέσεων του αρχικού δικτύου.

Στην πρώτη προσέγγιση RTSbinthr ακολουθείται η ίδια μέθοδο με την RTSweight με την αλλαγή στο βήμα 3 ότι δεν δημιουργείται ο πίνακας βαρών W^* αλλά ο πίνακας γειτνίασης A^* από τον πίνακα συσχετίσεων R^* των υποκατάστατων χρονοσειρών. Η μετατροπή του πίνακα R^* στον A^* γίνεται με τον ίδιο τρόπο όπως στο αρχικό δίκτυο, δηλαδή με την χρήση κάποιο κατωφλίου ή με κάποιο τεστ στατιστικής σημαντικότητας για τα στοιχεία $r^*_{i,j}$ του πίνακα R^*.

Η μέθοδος αυτή δεν διατηρεί ούτε το συνολικό βαθμό του αρχικού δικτύου στα τυχαιοποιημένα δίκτυα ούτε και τον βαθμό κάθε κόμβου σε αυτά. [45]

4.5.2.2. Μη-σταθμισμένα τυχαιοποιημένα δίκτυα με διατήρηση του πλήθους των συνδέσεων του αρχικού δικτύου.

Η δεύτερη προσέγγιση RTSbindeg εισάγει ένα κατώφλι για το ποια στοιχεία $r^*_{i,j}$ του πίνακα R^* θα μετουσιωθούν σε πραγματικές συνδέσεις έτσι ώστε κάθε τυχαιοποιημένο δίκτυο να έχει τον ίδιο αριθμό συνδέσεων με το πραγματικό.

Για την εύρεση του κατωφλιού αρχικά διατάσσονται τα στοιχεία του άνω τριγωνικού μέρους του πίνακα συσχετίσεων των υποκατάστατων χρονοσειρών R^* σε φθίνουσα σειρά. Το κατώφλι είναι η τιμή του στοιχείου που η θέση διάταξής του είναι ίση με το πλήθος των συνδέσεων στο αρχικό δίκτυο.

Η μέθοδος αυτή διατηρεί το συνολικό βαθμό του αρχικού δικτύου στα τυχαιοποιημένα δίκτυα αλλά δεν διατηρεί τον βαθμό κάθε κόμβου. [45]

Κεφάλαιο 5: Αξιολόγηση ελέγχων για modularity.

Στο κεφάλαιο αυτό παρουσιάζεται μία σύγκριση των δύο διαφορετικών προσεγγίσεων παραγωγής τυχαιοποιημένων δικτύων. Η πρώτη προσέγγιση παράγει τυχαιοποιημένα δίκτυα τυχαιοποιώντας τις συνδέσεις του αρχικού δικτύου ενώ η δεύτερη προσέγγιση παράγει τυχαιοποιημένα δίκτυα δημιουργώντας υποκατάστατες χρονοσειρές και στη συνέχεια κατασκευάζεται από αυτές το τυχαιοποιημένο δίκτυο. Η σύγκριση γίνεται σε προσομοιωτικά συστήματα αλλά και σε μία εφαρμογή οικονομικών δεδομένων της Morgan Stanley Capital International (MSCI). Όλες οι εφαρμογές αναπτύχθηκαν στο προγραμματιστικό περιβάλλον MATLAB και χρησιμοποιήθηκαν συναρτήσεις του Community Detection Toolbox του Αθανάσιου Κεχαγιά [56], και συναρτήσεις που αναπτύχθηκαν από τον Κουγιουμτζή Δημήτριο [57].

5.1. Προσομοίωση σε γνωστά συστήματα.

Πρώτα θα παρουσιαστούν αναλυτικά δύο παραδείγματα για την ανάλυση σε ένα μη σταθμισμένο και σε ένα σταθμισμένο δίκτυο συσχέτισης και στη συνέχεια θα παρατεθούν τα αποτελέσματα του ελέγχου στατιστικής σημαντικότητας για τις μεθόδους τυχαιοποίησης.

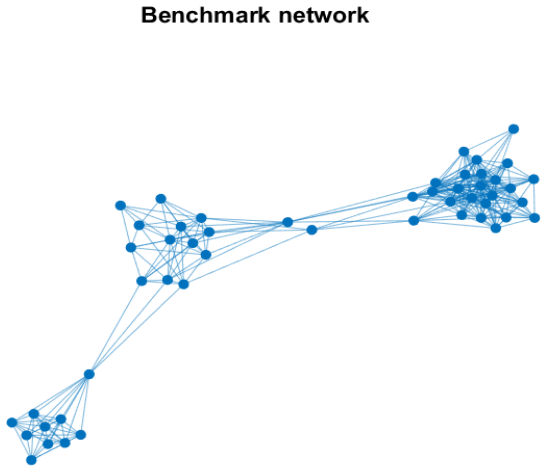
5.1.1. Δημιουργία γνωστού συστήματος.

Για την παραγωγή ενός γνωστού συστήματος 50 χρονοσειρών μεγέθους $T = 200$, δηλαδή ενός benchmark δικτύου, χρησιμοποιήθηκε ένα πολυμεταβλητό γραμμικό αυτοπαλινδρομούμενο μοντέλο τάξης 1, $VAR(1)$, σε $N = 50$ μεταβλητές, $x_t = Ax_{t-1} + e_t$. Στο μοντέλο δεν υπάρχουν όροι αλληλεξαρτήσεων, δηλαδή ο πίνακας A είναι διαγώνιος και τα στοιχεία της διαγώνιου που ορίζουν την ένταση της αυτοσυσχέτισης επιλέγονται τυχαία από την ομοιόμορφη κατανομή $U[0.01\,,0.9]$.

Το σφάλμα εισόδου ακολουθεί κανονική κατανομή $N(0\,,\Sigma_e)$. Για να εκχωρηθεί στο σύστημα η επιθυμητή δομή κοινοτήτων ο πίνακας συνδιασπορών Σ_e έχει μορφή μπλοκ διαγώνιου πίνακα ορίζοντας με αυτόν τον τρόπο το πλήθος των κοινοτήτων καθώς και τα μεγέθη τους. Τα στοιχεία εντός των τριών διαγώνιων μπλοκ επιλέγονται τυχαία από την ομοιόμορφη κατανομή $U[0.01\,,0.9]$ και έχουν μεγέθη $25, 15$ *και* 10 ενώ τα στοιχεία που δεν ανήκουν στα μπλοκ επιλέγονται τυχαία από την ομοιόμορφη κατανομή $U[0.01\,,0.2]$.

5.1.2. Παράδειγμα μη-σταθμισμένου δικτύου.

Από τις τιμές του συντελεστή συσχέτισης για κάθε ζεύγος των 50 χρονοσειρών δημιουργήθηκε ο πίνακας βαρών W θεωρώντας ως βάρος της σύνδεσης την απόλυτη τιμή του συντελεστή συσχέτισης. Ο πίνακας γειτνίασης του αρχικού δικτύου προέκυψε από τον πίνακα βαρών W θέτοντας ως κατώφλι αυθαίρετα το όριο 0.3. Το αρχικό δίκτυο φαίνεται στην Εικόνα 5.

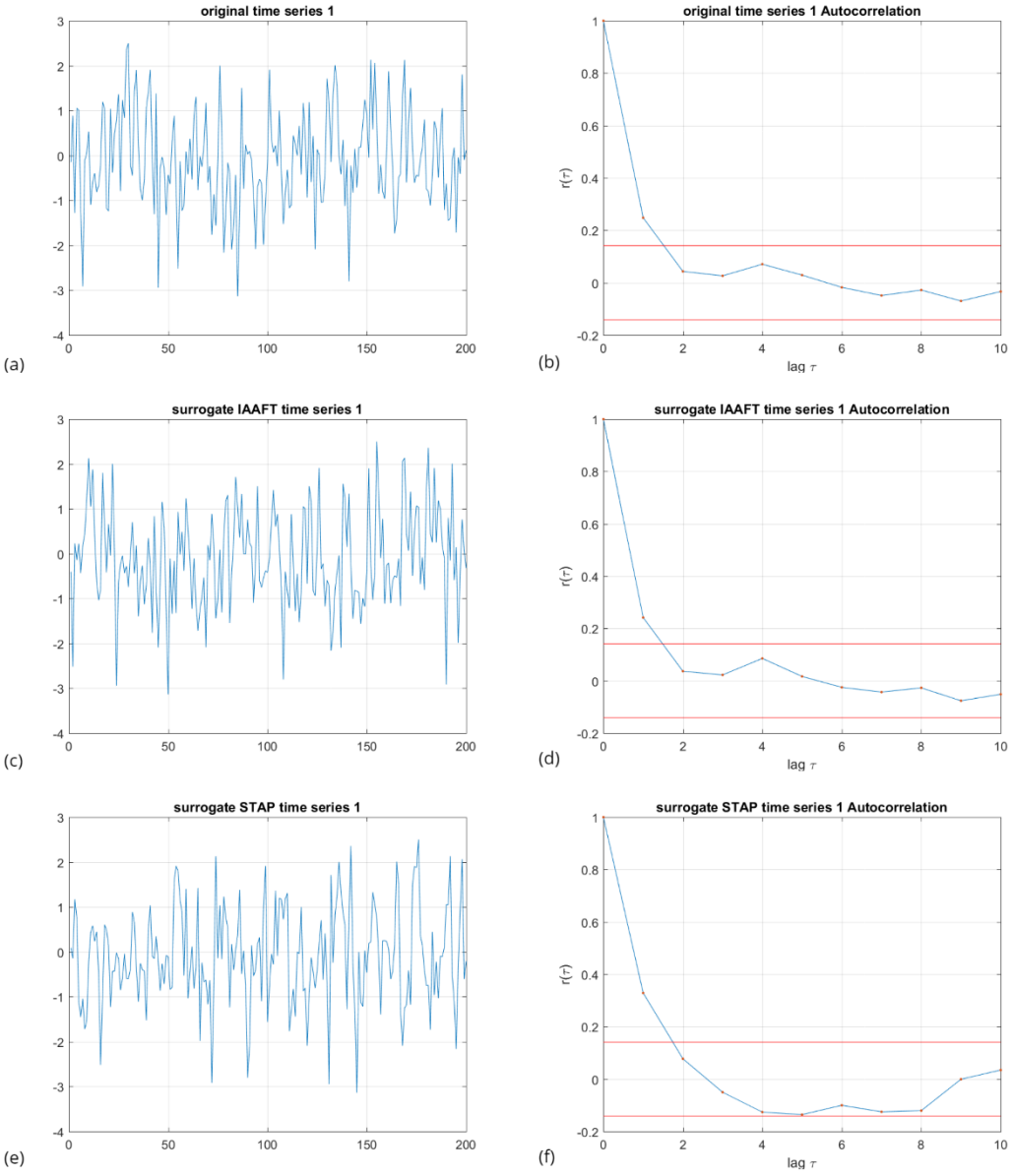


Εικόνα 5: Αρχικό μη σταθμισμένο δίκτυο συσχέτισης χρονοσειρών.

Για την δημιουργία των τυχαιοποιημένων δικτύων εφαρμόστηκε η μέθοδος RTSbindeg (παράγραφος 4.5.2.2.) που διατηρεί το συνολικό πλήθος των συνδέσεων του δικτύου, χρησιμοποιώντας τους αλγόριθμους δημιουργίας υποκατάστατων χρονοσειρών IAAFT και STAP οι οποίες αναφέρονται στη συνέχεια IAAFTbin και STAPbin αντίστοιχα καθώς και η μέθοδος Maslov (παράγραφος 4.2).

Στην Εικόνα 6, παρουσιάζεται η πρώτη χρονοσειρά από τις 50 που δημιουργήθηκαν αρχικά και η συνάρτηση αυτοσυσχέτισης της $(6a, 6b)$, οι υποκατάστατες χρονοσειρές και οι αυτοσυσχετίσεις τους που δημιουργήθηκαν με τον αλγόριθμο IAAFT $(6c, 6d)$, και με τον αλγόριθμο STAP $(6e, 6f)$. Παρατηρείται ότι η αυτοσυσχέτιση της υποκατάστατης χρονοσειράς που δημιουργήθηκε με τον αλγόριθμο IAAFT δεν διαφέρει καθόλου με την αρχική, ενώ η αυτοσυσχέτιση της υποκατάστατης χρονοσειράς που δημιουργήθηκε με τον αλγόριθμο STAP διαφέρει για χρονική

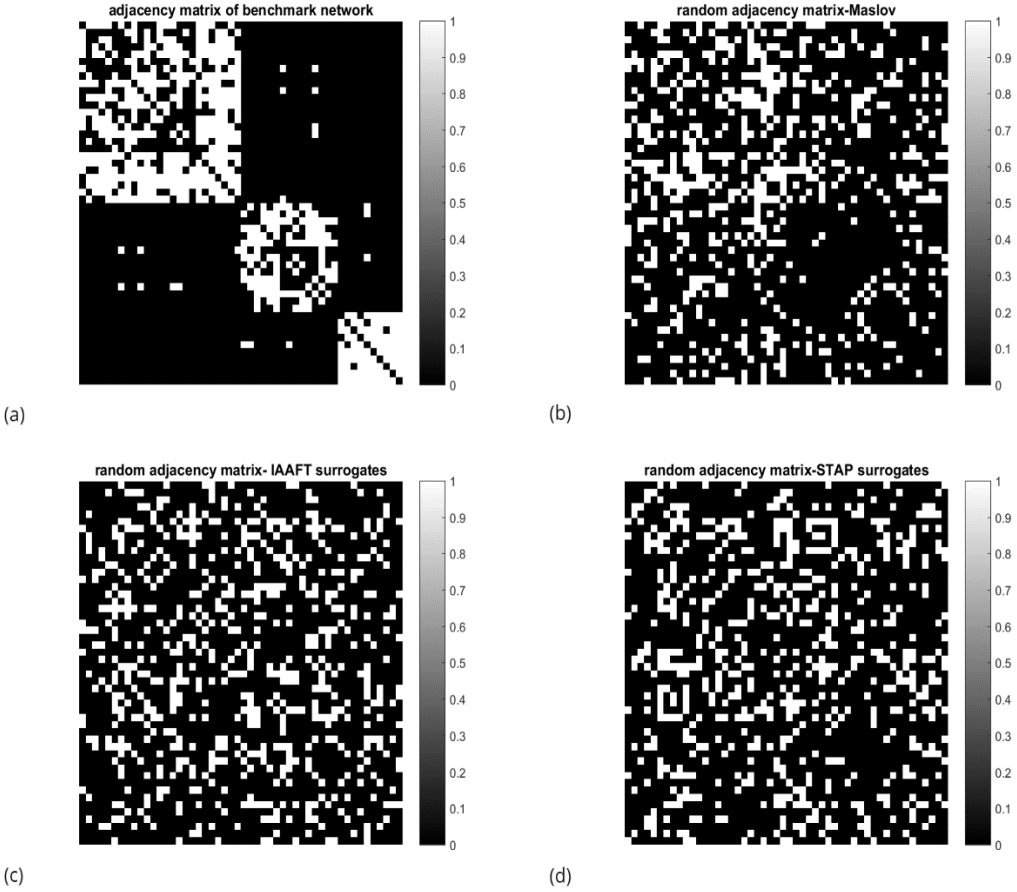
υστέρηση $\tau \geq 3$. Ωστόσο η διαφορά δεν είναι ουσιαστική καθώς είναι για αυτοσυσχέτιση στο επίπεδο του μηδενός όπως φαίνεται και από τις κόκκινες γραμμές.



Εικόνα 6:Η πρώτη χρονοσειρά (a), η αυτοσυσχέτισή της (b), υποκατάστατη χρονοσειρά με την μέθοδο IAAFT (c) και η αυτοσυσχέτισή της (d), υποκατάστατη χρονοσειρά με την μέθοδο STAP (e) και η αυτοσυσχέτισή της (f).

Στην Εικόνα 7 παρουσιάζονται οι πίνακες γειτνίασης του πραγματικού δικτύου $(7a)$ και των τυχαιοποιημένων δικτύων με τις μεθόδους Maslov $(7b)$, IAAFTbin $(7c)$ και STAPbin $(7d)$. Παρατηρείται ότι ο πίνακας γειτνίασης του αρχικού δικτύου έχει την μπλοκ διαγώνια μορφή του πίνακα συνδιασπορών Σ_e, και επομένως το δίκτυο παρουσιάζει ισχυρή δομή κοινοτήτων αφού σχεδόν όλο το πλήθος των συνδέσεων

βρίσκεται μέσα στα διαγώνια μπλοκ, δηλαδή μεταξύ κόμβων που ανήκουν στην ίδια κοινότητα. Αυτό μπορεί να διαπιστωθεί εύκολα ελέγχοντας και την παράμετρο μίξης, δηλαδή το ποσοστό των συνδέσεων που ενώνουν κόμβους του δικτύου που ανήκουν σε διαφορετικές κοινότητες επί του συνόλου τω συνδέσεων και στην συγκεκριμένη περίπτωση η τιμή της παραμέτρου είναι 0.03.

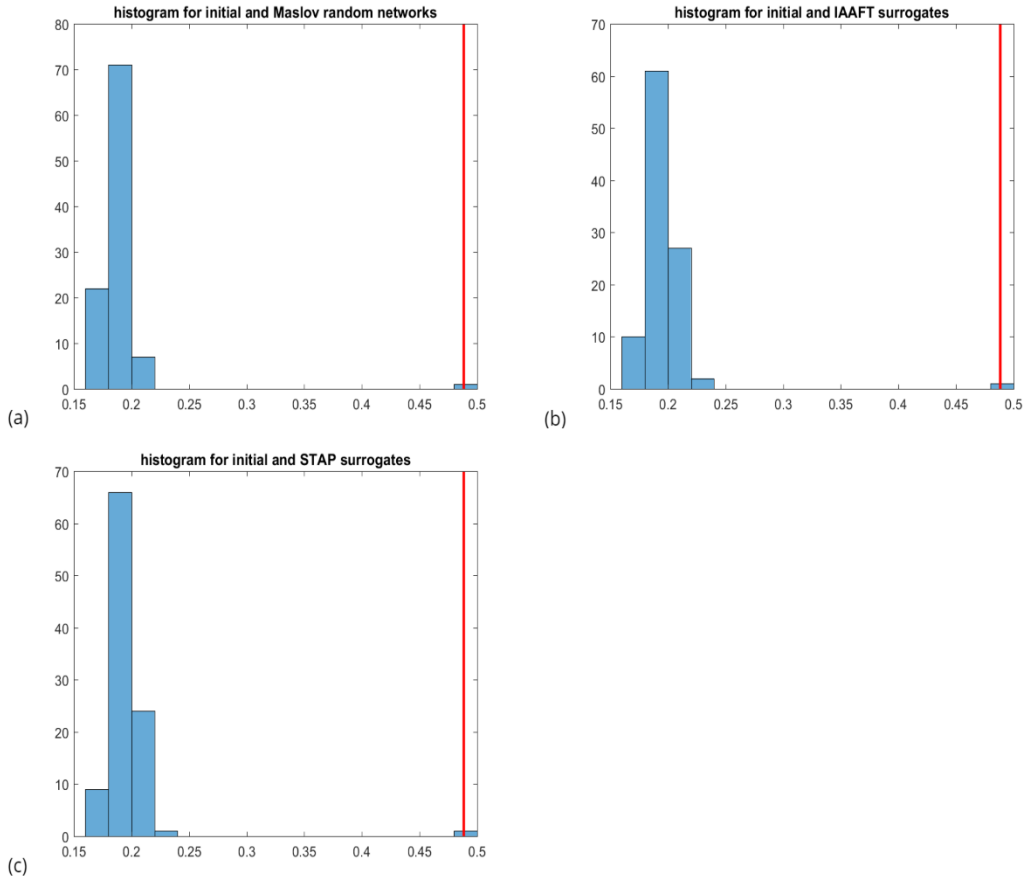


Εικόνα 7: Πίνακας γειτνίασης του αρχικού δικτύου (a), τυχαιοποιημένου δικτύου Maslov (b), IAAFTbin (c) και STAPbin (d).

Για κάθε μία από τις τρεις μεθόδους δημιουργήθηκαν $B = 100$ τυχαιοποιημένα δίκτυα. Ως στατιστικό του ελέγχου υπόθεσης με μηδενική υπόθεση H_0 ότι το αρχικό είναι τυχαίο, δηλαδή δεν παρουσιάζει δομή κοινοτήτων θεωρήθηκε το modularity και υπολογίστηκε η $\mathrm{p-τιμή} = 1 - \frac{r_0 - 0.326}{B+1+0.348}$ του μονόπλευρου ελέγχου για επίπεδο σημαντικότητας 0.05. Για τον υπολογισμό της τιμής modularity και την εύρεση κοινοτήτων χρησιμοποιήθηκε ο αλγόριθμος Louvain (παράγραφος 2.5.1).

Τα γραφήματα της Εικόνας 8 δείχνουν τις τιμές του modularity για το αρχικό δίκτυο (κόκκινη γραμμή) και τα 100 τυχαιοποιημένα δίκτυα με την μέθοδο Maslov

$(8a)$, IAAFTbin $(8b)$ και STAPbin $(8c)$. Παρουσιάζονται τα ιστογράμματα για κάθε μία μέθοδο τυχαιοποίησης και διαπιστώνεται αν η τιμή modularity του αρχικού δικτύου που χρησιμοποιείται ως στατιστικό του ελέγχου ανήκει μέσα στην κατανομή των τιμών των αντίστοιχων τυχαιοποιημένων δικτύων. Η p-τιμή και για τις τρεις μεθόδους είναι 0.0067, καθώς η τιμή modularity του αρχικού δικτύου είναι μεγαλύτερη από τις τιμές modularity όλων των τυχαιοποιημένων δικτύων και επομένως η μηδενική υπόθεση H_0 ότι το δίκτυο είναι τυχαίο και δεν παρουσιάζει δομή κοινοτήτων απορρίπτεται και γίνεται δεκτή η εναλλακτική υπόθεση H_1 ότι το δίκτυο παρουσιάζει δομή κοινοτήτων.



Εικόνα 8: Ιστόγραμμα της εμπειρικής κατανομής του modularity με τις μεθόδους Maslov (a), IAAFTbin (b) και STAPbin (c).

Ως στατιστικό του ελέγχου υπόθεσης με μηδενική υπόθεση H_0 ότι η κοινότητα είναι τυχαία θεωρήθηκε το πλήθος των συνδέσεων των κοινοτήτων στα τυχαιοποιημένα δίκτυα. Σε κάθε ένα από τα τυχαιοποιημένα δίκτυα χρησιμοποιήθηκε ο αλγόριθμος K-means ώστε να βρεθεί η βέλτιστη διαμέριση για το τυχαιοποιημένο δίκτυο με τόσες κοινότητες όσες βρέθηκαν με τον αλγόριθμο Louvain στο αρχικό δίκτυο. Στη συνέχεια η p-τιμή υπολογίστηκε ως η πιθανότητα να βρεθεί στο τυχαιοποιημένο δίκτυο μία κοινότητα με μεγαλύτερο πλήθος εσωτερικών συνδέσεων από το πλήθος των συνδέσεων του αρχικού δικτύου για επίπεδο σημαντικότητας 0.05. Σημειώνεται ότι η κοινότητα με το μεγαλύτερο πλήθος κόμβων στο πραγματικό δίκτυο συγκρίνεται με την κοινότητα με το μεγαλύτερο πλήθος κόμβων στο τυχαιοποιημένο δίκτυο και ου το καθεξής.

Στον Πίνακα 1 παρουσιάζονται οι p-τιμές για τρεις κοινότητες του αρχικού δικτύου σύμφωνα με τις τρεις μεθόδους τυχαιοποίησης δικτύων. Για επίπεδο σημαντικότητας 0.05 σύμφωνα με την μέθοδο Maslov η μηδενική υπόθεση H_0 ότι η κοινότητα είναι τυχαία απορρίπτεται και για τις τρεις κοινότητες. Οι μέθοδοι IAAFTbin και STAPbin απορρίπτουν την μηδενική υπόθεση για την πρώτη και την τρίτη κοινότητα, ενώ για την δεύτερη κοινότητα η μηδενική υπόθεση γίνεται δεκτή δηλαδή θεωρείται τυχαία.

Πίνακας 1: p-τιμές για τις τρεις κοινότητες του αρχικού μη σταθμισμένου δικτύου.

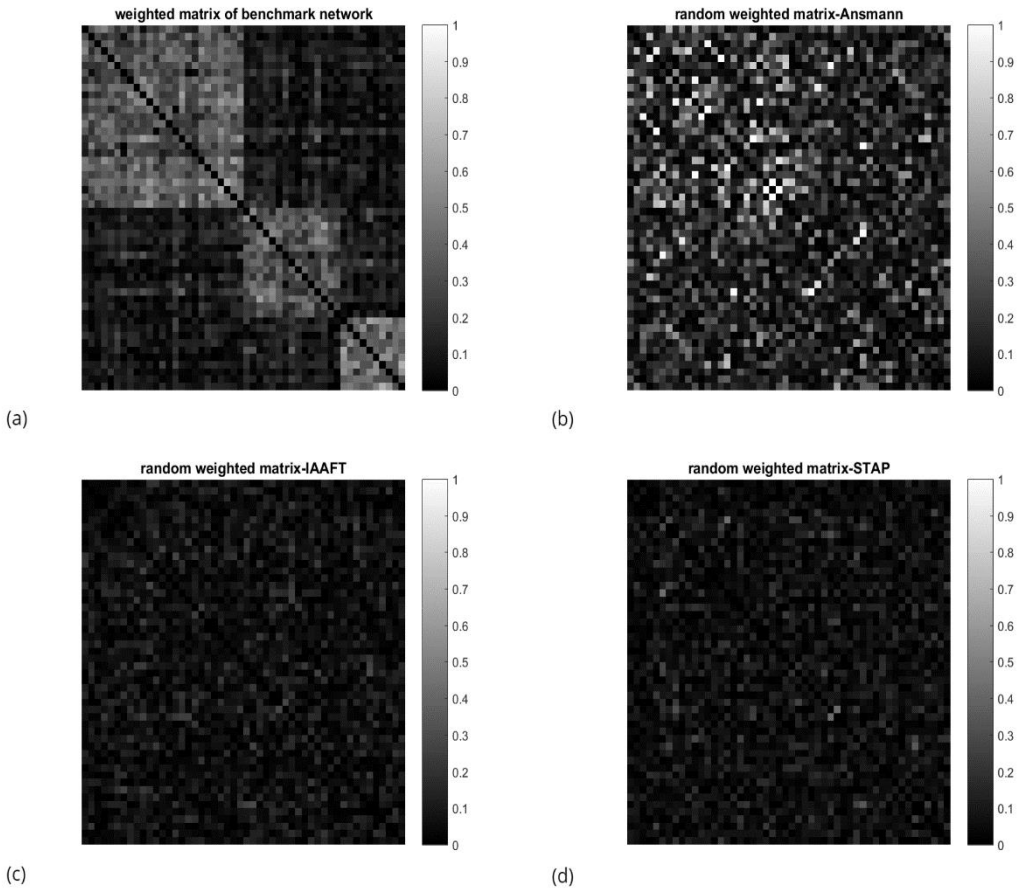
	Πλήθος κόμβων κοινότητας	Maslov	IAAFTbin	STAPbin
Κοινότητα 1	25	0.0067	0.03	0.01
Κοινότητα 2	15	0.04	0.06	0.06
Κοινότητα 3	10	0.0067	0.0067	0.0067

5.1.3. Παράδειγμα σταθμισμένου δικτύου.

Στην ανάλυση του σταθμισμένου δικτύου χρησιμοποιήθηκε ο πίνακας βαρών W. Για την δημιουργία των τυχαιοποιημένων σταθμισμένων δικτύων εφαρμόστηκε η μέθοδος RTSweight (παράγραφος 4.5.1.) η οποία δεν διατηρεί τη συνολική ισχύ του δικτύου, χρησιμοποιώντας τους αλγόριθμους δημιουργίας υποκατάστατων χρονοσειρών IAAFT και STAP οι οποίες αναφέρονται στη συνέχεια IAAFTwei και STAPwei αντίστοιχα, καθώς και η μέθοδος Ansmann (παράγραφος 3.3) η οποία διατηρεί την συνολική ισχύ του δικτύου αλλά και κάθε κόμβου όμως όπως αναφέρθηκε νωρίτερα δεν είναι κατάλληλη για τυχαιοποίηση δικτύων συσχέτισης διότι υπάρχει η περίπτωση να δημιουργήσει συνδέσεις με βάρος μεγαλύτερο της μονάδας.

Στην Εικόνα 9 παρουσιάζονται οι πίνακες βαρών του πραγματικού δικτύου $(9a)$ και των τυχαιοποιημένων δικτύων με τις μεθόδους Ansmann $(9b)$, IAAFTwei $(9c)$ και STAPwei $(9d)$. Ο πίνακας βαρών του αρχικού δικτύου έχει την μπλοκ διαγώνια μορφή του πίνακα συνδιασπορών Σ_e, και επομένως το δίκτυο παρουσιάζει δομή κοινοτήτων, πιο ασθενή βέβαια από το μη-σταθμισμένο δίκτυο αφού η παράμετρος μίξης είναι 0.34.

Επίσης διακρίνεται διαφορά στην συνολική ισχύ ανάμεσα στον πίνακα βαρών του αρχικού δικτύου και στον τυχαίο πίνακα βαρών που προέκυψε με την μέθοδο Ansmann με τους τυχαίους πίνακες βαρών που δημιουργήθηκαν με τις μεθόδους IAAFTwei και STAPwei. Πιο συγκεκριμένα ο αρχικός πίνακας βαρών έχει συνολική ισχύ 236.95, η οποία διατηρείται με την μέθοδο τυχαιοποίησης Ansmann. Οι πίνακες των εικόνων $(9c)$ και $(9d)$ έχουν συνολική ισχύ 89.64 και 85.03 αντίστοιχα. Για τα 100 τυχαιοποιημένα δίκτυα που δημιουργήθηκαν για τον έλεγχο της μηδενική υπόθεσης με την μέθοδο IAAFTwei η μέση τιμή της συνολικής ισχύς είναι 90.5 ενώ με τη μέθοδο STAPwei είναι 86.16.

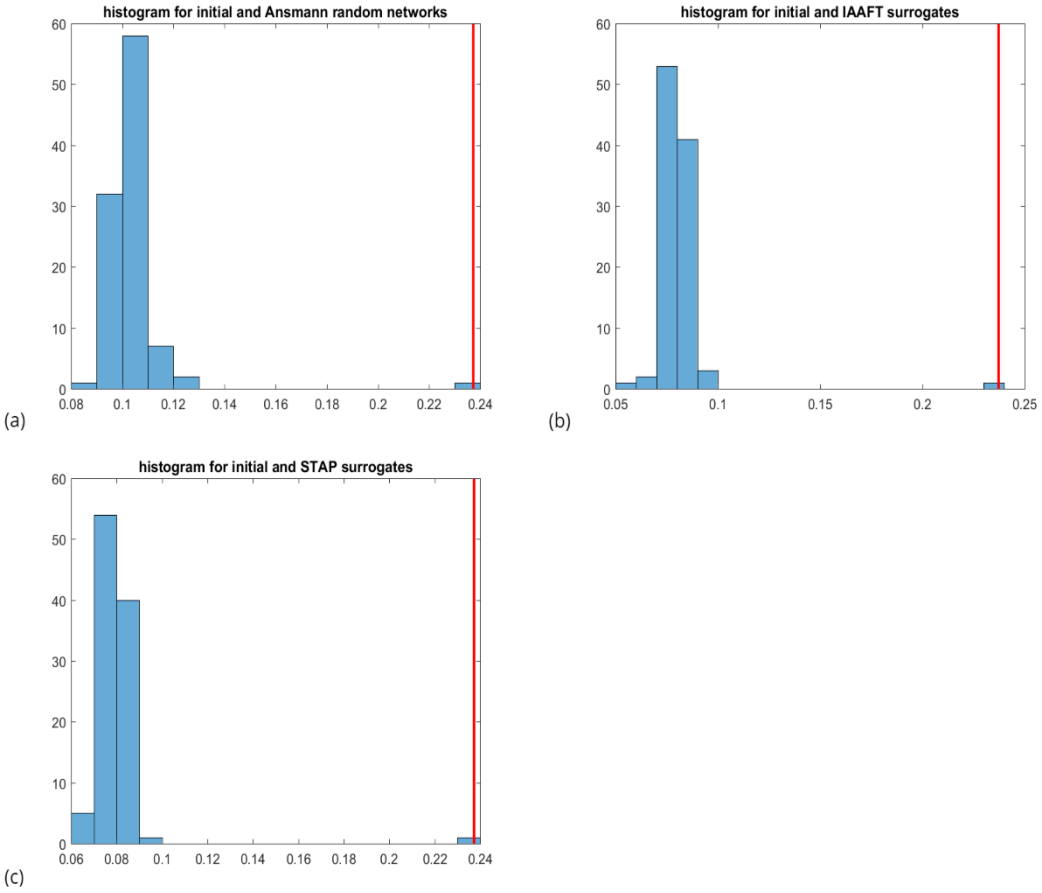


Εικόνα 9: Πίνακας βαρών για το αρχικό δίκτυο (a) και τα τυχαιοποιημένα με μέθοδο Ansmann (b), IAAFTwei (c) και STAPwei (d).

Η διαδικασία που ακολουθήθηκε για τον έλεγχο της μηδενικής υπόθεσης H_0 ότι το αρχικό είναι τυχαίο, δηλαδή δεν παρουσιάζει δομή κοινοτήτων, αλλά και για την κάθε κοινότητα ξεχωριστά με μηδενική υπόθεση H_0 ότι είναι τυχαία είναι όμοια με το παράδειγμα για το μη-σταθμισμένο δίκτυο (παράγραφος 5.1.2).

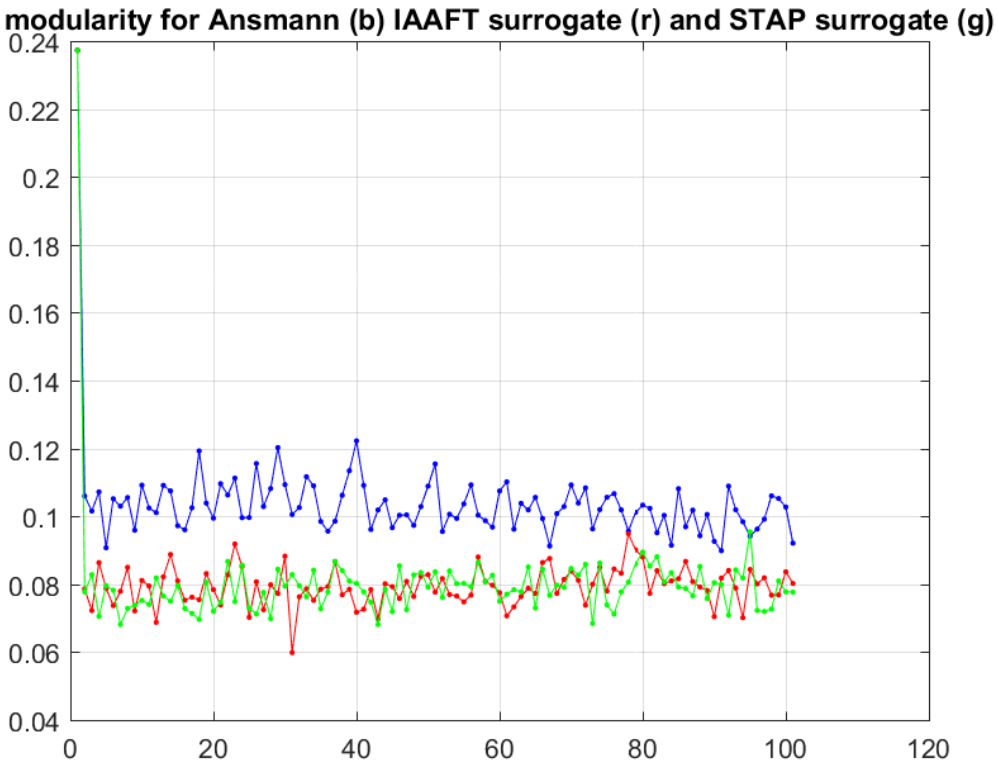
Τα γραφήματα της Εικόνας 10 δείχνουν τις τιμές του modularity για το αρχικό δίκτυο (κόκκινη γραμμή) και τα 100 τυχαιοποιημένα δίκτυα με την μέθοδο Ansmann ($10a$), IAAFTwei ($10b$) και STAPwei ($10c$). Παρουσιάζονται τα ιστογράμματα για κάθε μία μέθοδο τυχαιοποίησης και διαπιστώνεται αν η τιμή modularity του αρχικού δικτύου που χρησιμοποιείται ως στατιστικό του ελέγχου ανήκει μέσα στην κατανομή των τιμών των αντίστοιχων τυχαιοποιημένων δικτύων. Η p-τιμή και για τις τρεις μεθόδους είναι 0.0067, καθώς η τιμή modularity του αρχικού δικτύου είναι μεγαλύτερη από τις τιμές modularity όλων των τυχαιοποιημένων δικτύων και επομένως η μηδενική

υπόθεση H_0 ότι το δίκτυο είναι τυχαίο και δεν παρουσιάζει δομή κοινοτήτων απορρίπτεται και γίνεται δεκτή η εναλλακτική υπόθεση H_1.



Εικόνα 10: Ιστόγραμμα της εμπειρικής κατανομής του modularity με τις μεθόδους Ansmann (a), IAAFTwei (b) και STAPwei (c).

Στην Εικόνα 11 παρουσιάζονται οι τιμές του στατιστικού του ελέγχου, δηλαδή του modularity για το αρχικό δίκτυο και για τα τυχαιοποιημένα δίκτυα σύμφωνα και με τις τρεις μεθόδους, με μπλε χρώμα για την μέθοδο Ansmann, κόκκινο για την μέθοδο IAAFTwei και πράσινο για την μέθοδο STAPwei. Παρατηρείται ότι η μέθοδος Ansmann παράγει συστηματικά τυχαιοποιημένα δίκτυα με μεγαλύτερες τιμές modularity σε σχέση με τις δύο άλλες μεθόδους. Αυτό οφείλεται στο ότι η μέθοδος Ansmann διατηρεί την συνολική ισχύ του αρχικού δικτύου ενώ τα τυχαιοποιημένα δίκτυα που δημιουργούνται με τις μεθόδους IAAFTwei και STAPwei έχουν αρκετά μικρότερη συνολική ισχύ.



Εικόνα 11: Τιμές modularity για τα τυχαιοποιημένα δίκτυα με Ansmann (μπλε), IAAFTwei (κόκκινο) και STAPwei (πράσινο).

Στον Πίνακα 2 παρουσιάζονται οι p-τιμές για τρεις κοινότητες του αρχικού δικτύου σύμφωνα με τις τρεις μεθόδους τυχαιοποίησης δικτύων. Για επίπεδο σημαντικότητας 0.05 και σύμφωνα με τις μεθόδους IAAFTwei και STAPwei και για τις τρεις κοινότητες η μηδενική υπόθεση H_0 ότι είναι τυχαίες απορρίπτεται. Σύμφωνα με την μέθοδο Ansmann η μηδενική υπόθεση απορρίπτεται για τις κοινότητες μεγέθους 25 και 15, ενώ για την κοινότητα μεγέθους 10 η μηδενική υπόθεση γίνεται δεκτή.

Πίνακας 2: p-τιμές για τις τρεις κοινότητες του αρχικού σταθμισμένου δικτύου.

	Πλήθος κόμβων κοινότητας	Ansmann	IAAFTwei	STAPwei
Κοινότητα 1	25	0.0067	0.0067	0.0067
Κοινότητα 2	15	0.04	0.0067	0.0067
Κοινότητα 3	10	0.07	0.0067	0.0067

5.1.4. Έλεγχος της ισχύς και του μεγέθους του ελέγχου υπόθεσης.

Για την αξιολόγηση του ελέγχου υπόθεσης είναι απαραίτητο να εξεταστούν η ισχύς και κυρίως το μέγεθος του. Η ισχύς του ελέγχου είναι το ποσοστό των αληθώς θετικών τεστ, δηλαδή το ποσοστό απόρριψης της μηδενικής υπόθεσης H_0 όταν αυτή είναι ψευδής. Επομένως οι χρονοσειρές που θα δημιουργηθούν πρέπει να είναι συσχετισμένες ώστε τα αρχικά δίκτυα να έχουν δομή κοινοτήτων. Αντίστοιχα το μέγεθος του ελέγχου είναι το ποσοστό των ψευδώς θετικών τεστ (σφάλμα τύπου I), δηλαδή το ποσοστό απόρριψης της μηδενική υπόθεσης H_0 όταν αυτή είναι αληθής. Άρα οι χρονοσειρές που θα δημιουργήσουν τα αρχικά δίκτυα σε αυτή την περίπτωση πρέπει να είναι ασυσχέτιστες και ο πίνακας συνδιασπορών να μην έχει διαγώνια μπλοκ μορφή.

Δημιουργήθηκαν πολυμεταβλητές χρονοσειρές μεγέθους $T = 200$, χρησιμοποιώντας ένα πολυμεταβλητό γραμμικό αυτοπαλινδρομούμενο μοντέλο τάξης 1, $VAR(1)$, σε $N = 50$ μεταβλητές, $x_t = Ax_{t-1} + e_t$, για διαφορετικές εντάσεις αυτοσυσχέτισης, διαφορετικό πλήθος κοινοτήτων και διαφορετικές μορφές του πίνακα συνδιασπορών Σ_e. Για κάθε συνδυασμό δημιουργήθηκαν 100 πραγματοποιήσεις πολυμεταβλητών χρονοσειρών και για κάθε μία από αυτές δημιουργήθηκαν $B = 100$ τυχαιοποιημένα δίκτυα για τις 6 μεθόδους, για μη-σταθμισμένες συνδέσεις Maslov, IAAFTbin και STAPbin και για σταθμισμένες συνδέσεις Ansmann, IAAFTwei και STAPwei. Για κάθε ένα B τυχαιοποιημένα δίκτυα της κάθε μεθόδου και για τα αρχικά δίκτυα (μη-σταθμισμένα και σταθμισμένα) υπολογίστηκε η τιμή modularity.

Θεωρώντας ως στατιστικό του ελέγχου το modularity υπολογίστηκε η p-τιμή του ελέγχου για κάθε μέθοδο καθώς και η απόφαση του ελέγχου για επίπεδο σημαντικότητας 0.05. αυτή η διαδικασία επαναλήφθηκε για κάθε μία από τις 100 πραγματοποιήσεις.

Πρέπει να σημειωθεί ότι δεν εκτιμήθηκε η ισχύς και το μέγεθος του ελέγχου υποθέσεως για κάθε κοινότητα διότι ο αλγόριθμος Louvain δεν έχει ως αποτέλεσμα τον ίδιο αριθμό κοινοτήτων για κάθε μία από τις 100 πραγματοποιήσεις πολυμεταβλητών χρονοσειρών ενός σεναρίου καθιστώντας επομένως αδύνατη τη σύγκριση.

Στους Πίνακες 3 και 4 παρουσιάζονται τα συγκεντρωτικά αποτελέσματα της πιθανότητας απόρριψης της μηδενικής υπόθεσης για τις έξι μεθόδους δημιουργίας

τυχαιοποιημένων δικτύων για κάθε συνδυασμό των πραγματοποιήσεων των πολυμεταβλητών χρονοσειρών για στάθμη σημαντικότητας $a = 0.05$. Ο πρώτος πίνακας εξετάζει σενάρια στα οποία έχουν εμφωλευτεί τρεις κοινότητες με μεγέθη $25, 15, 10$ ενώ ο δεύτερος δύο ισοπληθές κοινότητες. Στην πρώτη στήλη δηλώνεται η ομοιόμορφη κατανομή από την οποία επιλέχθηκαν με τυχαίο τρόπο τα στοιχεία του διαγώνιου πίνακα A και ορίζουν την ένταση της αυτοσυσχέτισης. Στη δεύτερη και τρίτη στήλη δηλώνονται οι ομοιόμορφες κατανομές από τις οποίες επιλέχθηκαν τυχαία τα στοιχεία των διαγώνιων μπλοκ και τα στοιχεία που δεν ανήκουν στα διαγώνια μπλοκ του πίνακα συνδιασπορών Σ_e αντίστοιχα. Στην τέταρτη στήλη δηλώνεται το αν η μηδενική υπόθεση H_0 ότι το αρχικό είναι τυχαίο και δεν παρουσιάζει δομή κοινοτήτων είναι ψευδής ή όχι δηλαδή αν η πιθανότητα απόρριψης της H_0 πρόκειται για την ισχύ ή το μέγεθος του ελέγχου αντίστοιχα. Η μηδενική υπόθεση θεωρήθηκε αληθής όταν για τα σενάρια ο αλγόριθμος Louvain απέτυχε να αναγνωρίσει τις εμφωλευμένες κοινότητες και η παράμετρος μίξης ξεπερνά την τιμή 0.5, δηλαδή οι περισσότερες συνδέσεις στο δίκτυο ενώνουν κόμβους διαφορετικών κοινοτήτων.

Πίνακας 3: Πιθανότητες απόρριψης της μηδενικής υπόθεσης T=200, N=50, πλήθος κοινοτήτων=3 (25-15-10), a=0.05

A	In block S	Out block S	H_0	Maslov	IAAFT bin	STAP bin	Ansmann	IAAFT wei	STAP wei
U[0.01 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.2]	Ψευδής	1	1	1	1	1	1
U[0.01 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.4]	Ψευδής	1	1	1	1	1	1
U[0.01 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.6	Αληθής	0.67	0.53	0.61	0.99	1	1
U[0.01 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.9]	Αληθής	0.54	0.32	0.32	0.98	1	1
U[0.01 , 0.9]	U[0.01 , 0.4]	U[0.01 , 0.4]	Αληθής	0.55	0.35	0.38	0.91	0.97	0.97
U[0.01 , 0.9]	U[0.01 , 0.3]	U[0.01 , 0.3]	Αληθής	0.55	0.29	0.37	0.91	0.97	0.97
U[0.01 , 0.9]	U[0.01 , 0.25]	U[0.01 , 0.25]	Αληθής	0.60	0.40	0.43	0.90	0.98	0.98
U[0.5 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.2]	Ψευδής	1	1	1	1	1	1
U[0.5 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.4]	Ψευδής	0.98	0.97	0.97	1	1	1
U[0.5 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.6	Αληθής	0.66	0.43	0.46	0.96	0.98	0.98
U[0.5 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.9]	Αληθής	0.66	0.38	0.43	0.85	0.93	0.96
U[0.5 , 0.9]	U[0.01 , 0.4]	U[0.01 , 0.4]	Αληθής	0.64	0.36	0.41	0.82	0.89	0.91
U[0.5 , 0.9]	U[0.01 , 0.3]	U[0.01 , 0.3]	Αληθής	0.41	0.20	0.22	0.63	0.77	0.79
U[0.5 , 0.9]	U[0.01 , 0.25]	U[0.01 , 0.25]	Αληθής	0.49	0.16	0.18	0.65	0.76	0.79

Πίνακας 4: Πιθανότητες απόρριψης της μηδενικής υπόθεσης T=200, N=50, πλήθος κοινοτήτων=2 (25-25), a=0.05

A	In blocks= S	Out block S	H_0	Maslov	IAAFT bin	STAP bin	Ansmann	IAAFT wei	STAP wei
U[0.01 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.2]	Ψευδής	1	1	1	1	1	1
U[0.01 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.4]	Ψευδής	1	1	1	1	1	1
U[0.01 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.6]	Αληθής	0.78	0.68	0.72	1	1	1
U[0.01 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.9]	Αληθής	0.58	0.29	0.32	0.97	0.99	1
U[0.01 , 0.9]	U[0.01 , 0.4]	U[0.01 , 0.4]	Αληθής	0.70	0.41	0.51	0.98	0.99	1
U[0.5 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.2]	Ψευδής	1	1	1	1	1	1
U[0.5 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.4]	Ψευδής	1	0.99	1	1	1	1
U[0.5 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.6]	Αληθής	0.73	0.50	0.55	0.95	0.97	0.97
U[0.5 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.9]	Αληθής	0.60	0.36	0.38	0.86	0.92	0.93
U[0.5 , 0.9]	U[0.01 , 0.4]	U[0.01 , 0.4]	Αληθής	0.56	0.25	0.30	0.80	0.93	0.94

Όσο αναφορά την ισχύ του ελέγχου υπόθεσης, δηλαδή όταν η μηδενική υπόθεση H_0 είναι ψευδής παρατηρείται ότι όλες οι μέθοδοι είτε για μη-σταθμισμένο δίκτυο είτε για σταθμισμένο δίκτυο έχουν πολύ υψηλή ισχύ και κρίνονται ότι αποδίδουν το ίδιο καλά. Δηλαδή το σφάλμα τύπου II είναι σχεδόν 0.

Όσο αναφορά το μέγεθος του ελέγχου υπόθεσης, δηλαδή όταν η μηδενική υπόθεση H_0 είναι αληθής παρατηρείται ότι το «πραγματικό μέγεθος» του ελέγχου απέχει αρκετά και δεν προσεγγίζει σε καμία περίπτωση το «κατ'όνομα» μέγεθος του ελέγχου $a = 0.05$.

Στην περίπτωση των μη-σταθμισμένων δικτύων οι μέθοδοι IAAFTbin και STAPbin υπερτερούν της μεθόδου Maslov όμως η εκτίμηση του μεγέθους του ελέγχου στην καλύτερη περίπτωση είναι τρεις έως τέσσερις φορές μεγαλύτερη από το επίπεδο σημαντικότητας $a = 0.05$, ακόμα για αρκετά αραιά δίκτυα.

Στην περίπτωση των σταθμισμένων δικτύων παρατηρείται ότι όλες οι μέθοδοι αποτυγχάνουν να πλησιάσουν το «κατ'όνομα» μέγεθος του ελέγχου για όλα τα σενάρια και αντίθετα οι πιθανότητες απόρριψης κοντά στη μονάδα, κάτι που υποδηλώνει συστηματική μεροληψία υπέρ της απόρριψης της μηδενικής υπόθεσης. Επίσης η μέθοδος Ansmann έχει μικρότερες πιθανότητες απόρριψης της μηδενικής υπόθεσης,

κάτι που είναι αναμενόμενο αφού οι μέθοδοι IAAFTwei και STAPwei δημιουργούν τυχαιοποιημένα δίκτυα με αρκετά μικρότερη συνολική ισχύ.

Παρατηρείται σε αρκετές περιπτώσεις ότι η πιθανότητα απόρριψης σύμφωνα με τις μεθόδους IAAFTbin και STAPbin είναι μικρότερη όταν η ένταση της αυτοσυσχέτισης αυξάνεται. Αυτό δικαιολογείται για τις μεθόδους που τυχαιοποιούν τις χρονοσειρές. Οι δύο προαναφερθείσες μέθοδοι διατηρούν την συνάρτηση αυτοσυσχέτισης ων αρχικών χρονοσειρών για τις υποκατάστατες. Επιπλέον είναι γνωστό ότι η υψηλή αυτοσυσχέτιση έχει ως αποτέλεσμα την δημιουργία ψευδώς ισχυρών συνδέσεων. Αν λάβουμε υπόψη και το μοτίβο μεταβατικότητας που χαρακτηρίζει τα δίκτυα συσχέτισης, δηλαδή ότι ισχυρή θετική συσχέτιση στο μη-άμεσο μονοπάτι $(X - Y - Z)$ υπαινίσσεται συσχέτιση στην άμεση σύνδεση $(X - Z)$. Τότε γίνεται κατανοητό ότι οι μέθοδοι IAAFTbin και STAPbin δημιουργούν τυχαιοποιημένα δίκτυα τα οποία διατηρούν εμμέσως το μοτίβο μεταβατικότητας που χαρακτηρίζει το αρχικό δίκτυο και δημιουργούν εγγενώς πιο δομημένα δίκτυα.

Στον Πίνακα 5 παρουσιάζονται οι πιθανότητες απόρριψης της μηδενικής υπόθεσης για τις έξι μεθόδους δημιουργίας τυχαιοποιημένων δικτύων για σενάρια πραγματοποίησης των πολυμεταβλητών χρονοσειρών στις οποίες τα στοιχεία του πίνακα συνδιασπορών (πλην της διαγώνιου) είναι σταθερά. Αυτά τα σενάρια δημιουργήθηκαν με σκοπό να μειωθούν οι αποκλίσεις στις τιμές των στοιχείων του πίνακα συνδιασπορών και κατ' επέκταση στον πίνακα συσχετίσεων.

Πίνακας 5 Πιθανότητες απόρριψης της μηδενικής υπόθεσης T=200, N=50, a=0.05 και σταθερές τιμές για τα στοιχεία του πίνακα συνδιασπορών.

A	S	H_0	Maslov	IAAFT bin	STAP bin	Ansmann	IAAFT wei	STAP wei
U[0.01 , 0.9]	0.2	Αληθής	0.17	0.04	0.04	1	1	1
U[0.01 , 0.9]	0.25	Αληθής	0.33	0.11	0.09	1	1	1
U[0.01 , 0.9]	0.3	Αληθής	0.60	0.30	0.34	0.94	0.98	0.98
U[0.5 , 0.9]	0.2	Αληθής	0.40	0.06	0.06	1	1	1
U[0.5 , 0.9]	0.25	Αληθής	0.45	0.09	0.10	0.96	0.98	0.98
U[0.5 , 0.9]	0.3	Αληθής	0.46	0.16	0.16	0.67	0.82	0.83

Για τις μη-σταθμισμένες μεθόδους τυχαιοποίησης δικτύων παρατηρείται ότι οι μέθοδοι IAAFTbin και STAPbin υπερτερούν της μεθόδου Maslov για όλα τα σενάρια και ειδικά για το σενάριο $A \in U[0.01\,, 0.9]$ και $S = 0.2$ έχουν ποσοστό απόρριψης μικρότερο από 5%. Οι σταθμισμένες μέθοδοι τυχαιοποίησης δικτύων αντίθετα έχουν ποσοστά απόρριψης της μηδενικής υπόθεσης πολύ μεγαλύτερα από την στάθμη σημαντικότητας $a = 0.05$.

Στο παράρτημα δίνονται οι πίνακες με τις πιθανότητες απόρριψης για επίπεδο σημαντικότητας $a = 0.01$.

5.1.5. Συμπεράσματα προσομοίωσης σε γνωστά συστήματα.

Τα αποτελέσματα της προσομοιωτικής μελέτης σε γνωστά συστήματα για την εκτίμηση της ισχύς και του μεγέθους του ελέγχου υπόθεσης αναδεικνύουν δύο βασικά συμπεράσματα.

Αρχικά την αποτυχία της μεθόδου Maslov που εφαρμόζεται απευθείας στο δίκτυο συσχέτισης και τυχαιοποιεί τις συνδέσεις του, να δημιουργήσει τυχαιοποιημένα δίκτυα που να έχουν ίδια χαρακτηριστικά με το αρχικό δίκτυο συσχέτισης. Αντίθετα οι μέθοδοι IAAFTbin και STAPbin που τυχαιοποιούν τις χρονοσειρές από τις οποίες δημιουργείται το δίκτυο συσχέτισης έχουν καλύτερα αποτελέσματα, ωστόσο η αποτελεσματικότητά τους επηρεάζεται από την πυκνότητα του δικτύου.

Επίσης αναδεικνύεται η ακαταλληλότητα και των τριών μεθόδων τυχαιοποίησης σταθμισμένου δικτύου Ansmann, IAAFTwei και STAPwei για την δημιουργία τυχαιοποιημένων δικτύων που να έχουν ίδια χαρακτηριστικά με το αρχικό δίκτυο.

5.2. Εφαρμογή σε πραγματικά δεδομένα.

Η εφαρμογή σε πραγματικά δεδομένα θα διερευνήσει την τυχαιότητα ως προς την δομή κοινοτήτων των δικτύων συσχέτισης που δημιουργούνται από οικονομικά δεδομένα.

5.2.1. Περιγραφή των δεδομένων.

Τα δεδομένα που χρησιμοποιούνται για την εφαρμογή είναι ο δείκτης σταθμισμένης χρηματιστηριακής αξίας της Morgan Stanley Capital International (MSCI) για ένα σύνολο 55 αγορών για την χρονική περίοδο 5 Μαρτίου 2004 έως 5

Μαρτίου 2009. Ο δείκτης υπολογίζεται με τη βοήθεια της αξίας ιδίων κεφαλαίων εταιρειών οι οποίες είναι αντιπροσωπευτικές της δομής της αγοράς. Οι αγορές βρίσκονται σε Αμερική, Ευρώπη, Ασία/ περιοχή Ειρηνικού ωκεανού και Αφρική. Τα δεδομένα αποτελούνται από 1305 τιμές για κάθε αγορά και αναφέρονται σε διάστημα 4 ετών εξαιρουμένων Σαββατοκύριακων και αργιών.

Οι 55 αγορές χωρίζονται σε τρεις κατηγορίες: στις αναπτυγμένες, στις αναδυόμενες και στις αναπτυσσόμενες. Οι αναπτυγμένες αγορές είναι : Αυστρία, Βέλγιο, Δανία, Φινλανδία, Γαλλία, Γερμανία, Ελλάδα, Ιρλανδία, Ιταλία, Ολλανδία, Νορβηγία, Πορτογαλία, Ισπανία, Σουηδία, Ελβετία, Ηνωμένο Βασίλειο, Η.Π.Α., Καναδάς, Αυστραλία, Χονγκ Κονγκ, Ιαπωνία, Νέα Ζηλανδία, Σιγκαπούρη. Οι αναδυόμενες αγορές είναι : Τσεχία, Ουγγαρία, Ισραήλ, Πολωνία, Ρωσία, Τουρκία, Αργεντινή, Βραζιλία, Χιλή, Κολομβία, Μεξικό, Περού, Αίγυπτος, Μαρόκο, Νότιος Αφρική, Κίνα, Ινδία, Ινδονησία, Κορέα, Μαλαισία, Πακιστάν, Φιλιππίνες, Ταιβάν, Ταϊλάνδη. Οι αναπτυσσόμενες αγορές είναι : Κροατία, Σλοβενία, Μαυρίκιος, Ιορδανία, Κένυα, Λίβανος, Νιγηρία, Σρι Λάνκα.

5.2.2. Περιγραφή της εφαρμογής.

Για την ανάλυση των δεδομένων υπολογίστηκαν αρχικά οι λογαριθμικές διαφορές του δείκτη σταθμισμένης χρηματιστηριακής αξίας για κάθε αγορά, προκειμένου να απαλειφθεί η τάση και η εποχικότητα και οι χρονοσειρές να γίνουν στάσιμες. Επομένως η τιμή των αποδόσεων Y_t υπολογίζεται $Y_t = \ln X_t - \ln X_{t-1}$, όπου X_t η τιμή του του δείκτη σταθμισμένης χρηματιστηριακής αξίας την χρονική στιγμή t.

Λαμβάνοντας υπόψη τα αποτελέσματα της προσομοίωσης σε γνωστά συστήματα για το μέγεθος του ελέγχου (παράγραφος 5.1.4) και την καθολική αποτυχία των μεθόδων τυχαιοποίησης σταθμισμένων δικτύων η εφαρμογή πραγματοποιήθηκε μόνο για την περίπτωση μη-σταθμισμένου δικτύου. Επομένως δημιουργήθηκαν τυχαιοποιημένα δίκτυα με τις μεθόδους Maslov, IAAFTbin, και STAPbin.

Ως κόμβοι του δικτύου θεωρήθηκαν οι 55 αγορές και οι συνδέσεις του μη-σταθμισμένου δικτύου πραγματοποιούνται με την προϋπόθεση ότι οι τιμές του συντελεστή συσχέτισης Pearson είναι μεγαλύτερες κατά απόλυτη τιμή από το αυθαίρετο κατώφλι που τέθηκε και στην προσομοίωση ίσο με 0.3.

Η διαδικασία που ακολουθήθηκε είναι ίδια με τις προσομοιώσεις στην περίπτωση των γνωστών συστημάτων. Η μηδενική υπόθεση H_0 είναι ότι το δίκτυο είναι

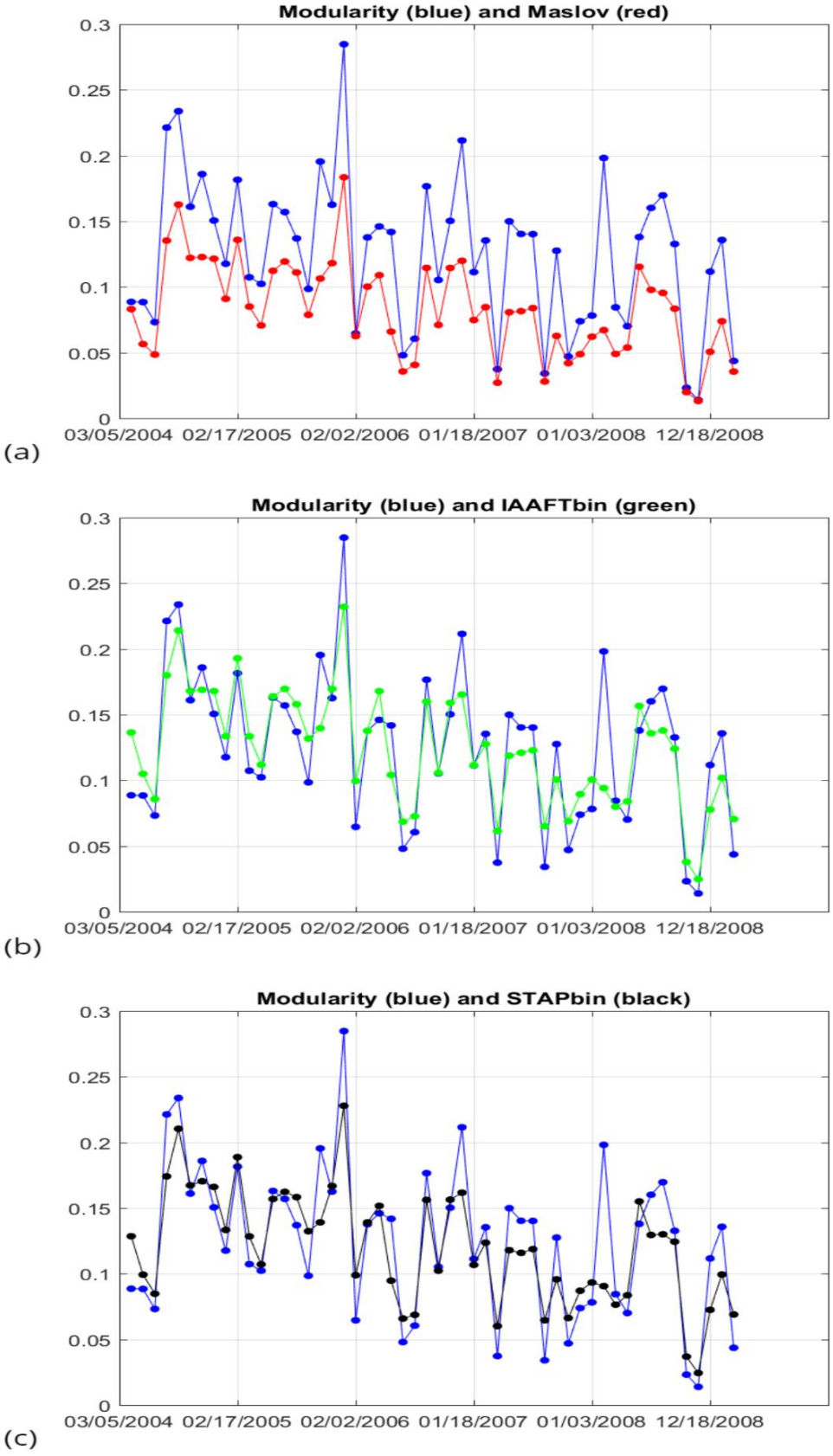
τυχαίο, δηλαδή δεν παρουσιάζει δομή κοινοτήτων και ως χαρακτηριστικό του ελέγχου θεωρήθηκε το modularity. Ως όρια για την τυχαιότητα ή όχι του δικτύου θεωρήθηκε η 96η τιμή modularity στην αύξουσα λίστα των τιμών των 100 τυχαιοποιημένων δικτύων και της τιμής του πραγματικού δικτύου και αντιστοιχεί σε $p - τιμή = 0.056$. Άρα όταν η γραφική παράσταση του modularity βρίσκεται μέσα στο όριο αυτό το δίκτυο θεωρείται τυχαίο ενώ σε διαφορετική περίπτωση θεωρείται πως παρουσιάζει στατιστικά σημαντική δομή κοινοτήτων.

Πραγματοποιήθηκαν δύο διαφορετικές αναλύσεις με βάση το χρονικό διάστημα που χωρίστηκε ολόκληρη η πολυμεταβλητή χρονοσειρά διάρκειας πέντε ετών και το αν τα διάστημα πάρθηκαν με επικάλυψη του προηγούμενου διαστήματος ή όχι. Στην πρώτη περίπτωση η πολυμεταβλητή χρονοσειρά χωρίστηκε σε 52 διαστήματα μήκους $T = 25$ ημερών, χωρίς επικάλυψη του προηγούμενου διαστήματος. Στην δεύτερη περίπτωση η πολυμεταβλητή χρονοσειρά χωρίστηκε σε 25 διαστήματα μήκους $T = 100$ ημερών με επικάλυψη του μισού προηγούμενου διαστήματος.

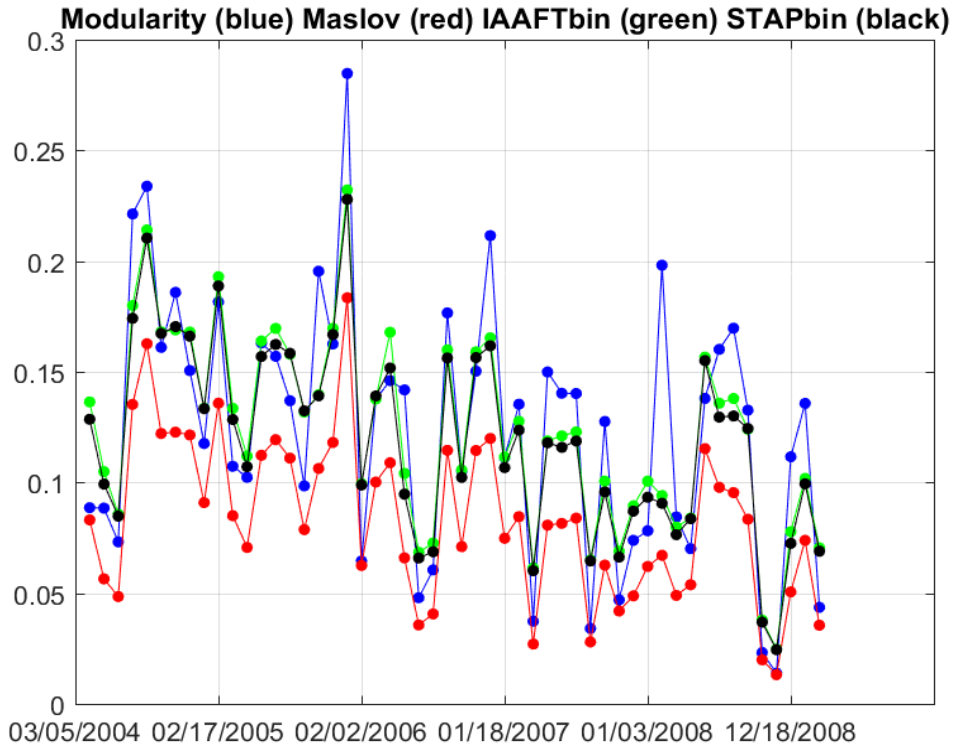
5.2.3. Αποτελέσματα εφαρμογής σε πραγματικά δεδομένα.

Στη συνέχεια παρουσιάζεται το modularity των πραγματικών δικτύων και τα όρια της τυχαιότητας σύμφωνα με κάθε μέθοδο, δηλαδή η 96η τιμή modularity στην αύξουσα λίστα των τιμών των 100 τυχαιοποιημένων δικτύων και της τιμής του πραγματικού δικτύου. Όταν η γραφική παράσταση του modularity των πραγματικών δικτύων είναι ίση ή βρίσκεται κάτω από τα όρια τυχαιότητας τότε το δίκτυο θεωρείται τυχαίο, ενώ σε αντίθετη περίπτωση παρουσιάζει στατιστικά σημαντική δομή κοινοτήτων.

Πρώτα παρουσιάζονται τα αποτελέσματα για τα διαστήματα μήκους $T = 25$, χωρίς επικάλυψη του προηγούμενου διαστήματος (Εικόνες 12 και 13) Με μπλε χρώμα παρουσιάζονται οι τιμές modularity των πραγματικών δικτύων, με κόκκινο χρώμα τα όρια τυχαιότητας σύμφωνα με την μέθοδο Maslov (Εικόνα $12a$), με πράσινο χρώμα τα όρια για την μέθοδο IAAFTbin (Εικόνα $12b$) και αντίστοιχα με μαύρο χρώμα για την μέθοδο STAPbin. (Εικόνα $12c$). Τέλος στην Εικόνα 13 παρουσιάζονται τα αποτλέσματα και με τις τρεις μεθόδους ταυτόχρονα.



Εικόνα 12: Οι τιμές modularity για τα πραγματικά δίκτυα, T=25 (μπλε) και τα όρια τυχαιότητας Maslov (κόκκινο), IAAFTbin (πράσινο), STAPbin (μαύρο).

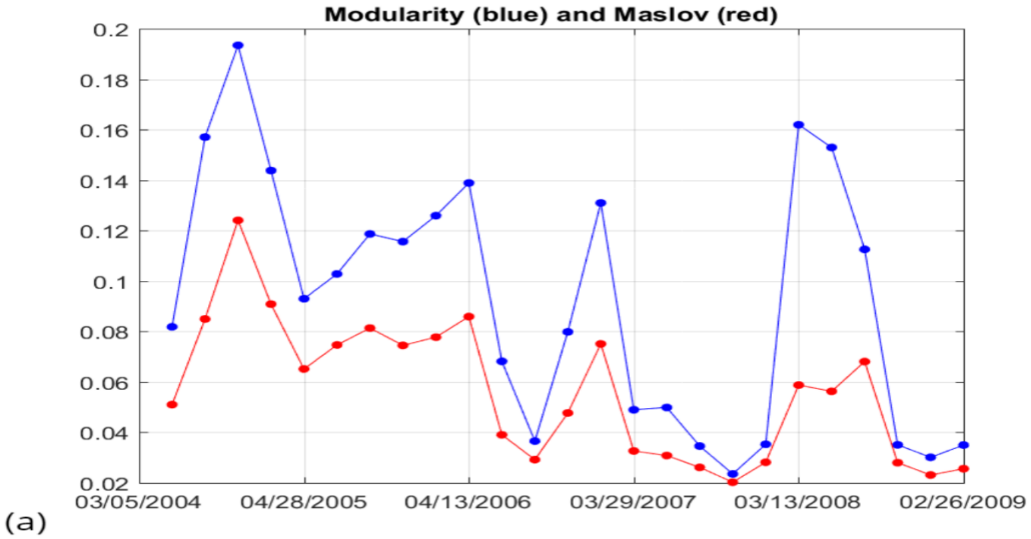


Εικόνα 13: Οι τιμές modularity και τα όρια τυχαιότητας με κάθε μέθοδο T=25.

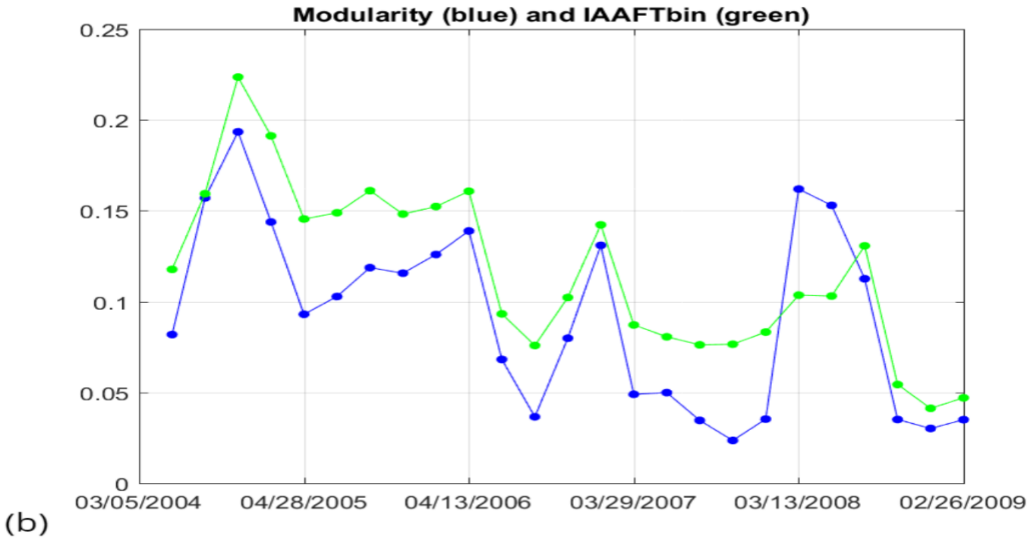
Παρατηρείται ότι με την μέθοδο Maslov η μηδενική υπόθεση ότι το δίκτυο είναι τυχαίο απορρίπτεται για όλα τα διαστήματα. Όσο αναφορά τις μεθόδους IAAFTbin και STAPbin δίνουν τα ίδια αποτελέσματα για την απόρριψή ή την αποδοχή της μηδενικής υπόθεσης, ωστόσο παρατηρείται ότι η μέθοδος STAPbin (μαύρη γραμμή) δίνει πιο συντηρητικά όρια τυχαιότητας σε σχέση με την μέθοδο IAAFTbin (πράσινη γραμμή).

Επίσης παρατηρείται από τα τέλη του έτους 2005 και μετά αρχίζουν και δημιουργούνται πιο συχνά δίκτυα με στατιστικά σημαντική δομή κοινοτήτων. Πιο συγκεκριμένα ο λόγος των μη-τυχαίων δικτύων προς το σύνολο τους μέχρι τα τέλη του 2005 είναι $\frac{4}{16} = 0.25$ και $\frac{5}{16} = 0.31$ σύμφωνα με την μέθοδο IAAFTbin και STAPbin αντίστοιχα. Ενώ για το διάστημα από τα τέλη του 2005 μέχρι το 2008 είναι $\frac{16}{33} = 0.48$ και $\frac{18}{33} = 0.54$. Δηλαδή παρουσιάζεται μία αύξηση της τάξης του 92% των μη τυχαίων δικτύων με την μέθοδο IAAFTbin και του 74% για την μέθοδο STAPbin αντίστοιχα.

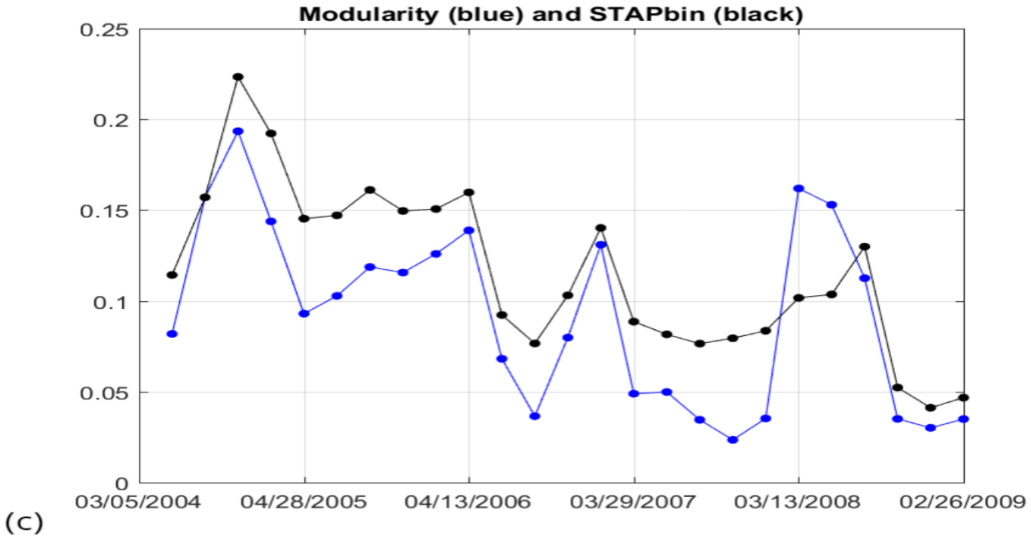
Στις Εικόνες 14 και 15 παρουσιάζονται τα αποτελέσματα για τα διαστήματα μήκους $T = 100$, με επικάλυψη του μισού προηγούμενου διαστήματος. Η χρωματική αντιστοιχία για κάθε μέθοδο παραμένει η ίδια.



(a)

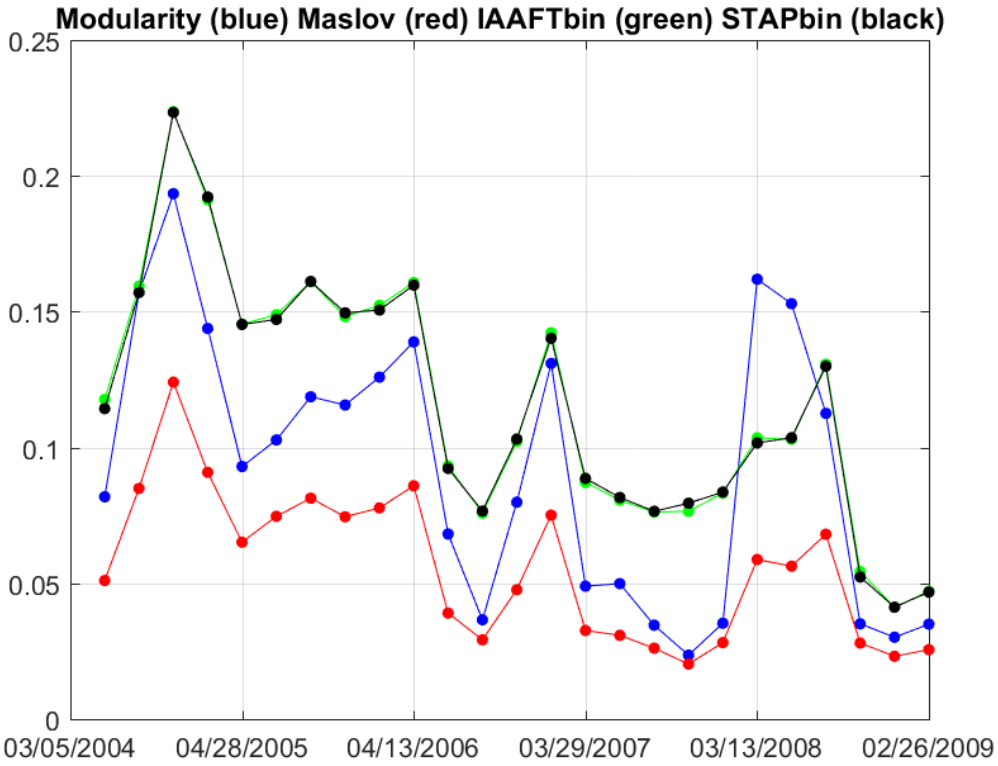


(b)



(c)

Εικόνα 14: Οι τιμές modularity για τα πραγματικά δίκτυα, T=100 με επικάλυψη του μισού προηγούμενου διαστήματος (μπλε) και τα όρια τυχαιότητας Maslov (κόκκινο), IAAFTbin (πράσινο), STAPbin (μαύρο).



Εικόνα 15: Οι τιμές modularity και τα όρια τυχαιότητας με κάθε μέθοδο T=100 με επικάλυψη του μισού προηγούμενου διαστήματος.

Παρατηρείται ότι και σε αυτήν την περίπτωση σύμφωνα με την μέθοδο Maslov τα δίκτυα συσχέτισης θεωρούνται ως μη-τυχαία. Οι μέθοδοι IAAFTbin και STAPbin και πάλι παρόμοια αποτελέσματα.

Επίσης παρατηρείται ότι σύμφωνα με τις μεθόδους IAAFTbin και STAPbin από την αρχή του έτους 2008 εμφανίζονται δίκτυα με στατιστικά σημαντική δομή κοινοτήτων για πρώτη φορά. Αυτή η περίοδος συμπίπτει με την εκδήλωση της παγκόσμιας χρηματοπιστωτικής κρίσης οπότε αυτό το γεγονός είναι σίγουρα μία πιθανή εξήγηση για την ξαφνική αλλαγή της δομής των δικτύων, κάτι που η μέθοδος Maslov αδυνατεί να αναγνωρίσει.

Κεφάλαιο 6: Συμπεράσματα διπλωματικής

Η παρούσα διπλωματική εργασία είχε ως αντικείμενο μελέτης την σύγκριση δύο διαφορετικών προσεγγίσεων για την δημιουργία τυχαιοποιημένων δικτύων συσχέτισης από πολυμεταβλητές χρονοσειρές σε σχέση με την μηδενική υπόθεση H_0 ότι το δίκτυο είναι τυχαίο και δεν παρουσιάζει δομή κοινοτήτων. Η πρώτη προσέγγιση τυχαιοποιεί απευθείας στις συνδέσεις του δικτύου σε αντίθεση με την δεύτερη που δημιουργεί υποκατάστατες χρονοσειρές και στη συνέχεια σχηματίζει από αυτές το τυχαιοποιημένο δίκτυο.

Ως στατιστικό του ελέγχου χρησιμοποιήθηκε το modularity και η εμπειρική κατανομή του δημιουργήθηκε από τις τιμές του για τα τυχαιοποιημένα δίκτυα που δημιουργήθηκαν έξι διαφορετικές μεθόδους τυχαιοποίησης.

Η προσομοίωση σε γνωστά συστήματα χρονοσειρών εκτίμησε τόσο την ισχύ όσο και το μέγεθος του ελέγχου υπόθεσης. Όσο αναφορά την ισχύ του ελέγχου τα αποτελέσματα είναι ικανοποιητικά για όλες τις μεθόδους σε μη-σταθμισμένα και σταθμισμένα δίκτυα. Αντίθετα τα αποτελέσματα της προσομοίωσης για το μέγεθος του ελέγχου αναδεικνύουν δύο βασικά συμπεράσματα. Αρχικά για την περίπτωση των μη-σταθμισμένων δικτύων την αναποτελεσματικότητα της μεθόδου Maslov που τυχαιοποιεί τις συνδέσεις του δικτύου καθώς είχε υψηλά ποσοστά απόρριψης της μηδενικής υπόθεσης ενώ ήταν αληθής, και την υπεροχή των μεθόδων IAAFTbin και STAPbin που τυχαιοποιούν τις χρονοσειρές παρόλο που σε λίγες περιπτώσεις το «πραγματικό μέγεθος» του ελέγχου προσέγγισε το «κατ' όνομα» μέγεθος του ελέγχου. Αφετέρου για την περίπτωση των σταθμισμένων δικτύων όλες οι μέθοδοι τυχαιοποίησης απέτυχαν να προσεγγίσουν το κατ' όνομα» μέγεθος του ελέγχου. Η αποτυχία των μεθόδων IAAFTwei και STAPwei οφείλεται στο ότι τα τυχαιοποιημένα δίκτυα που παράγουν παρουσιάζουν μεγάλη απώλεια στο συνολικό βάρος των συνδέσεων.

Επομένως ένας μελλοντικός σκοπός που δημιουργείται από την παρούσα διπλωματική εργασία είναι η δημιουργία ενός αλγορίθμου τυχαιοποίησης σταθμισμένων δικτύων για την μηδενική υπόθεση ότι το δίκτυο είναι τυχαίο και δεν παρουσιάζει δομή κοινοτήτων. Επίσης θα έχρηζε ενδιαφέροντος ο έλεγχος της μηδενικής υπόθεσης χρησιμοποιώντας ως στατιστικό κάποια άλλη ποιοτική συνάρτηση για την αξιολόγηση της διαμέρισης.

Η εφαρμογή στα πραγματικά δεδομένα απέδειξε την υπεροχή των μεθόδων IAAFTbin και STAPbin έναντι της μεθόδου Maslov σε ένα σύγχρονο οικονομικό πρόβλημα. Ωστόσο είναι ενδιαφέρον να απαντηθεί γιατί η αλλαγή της δομής των δικτύων συσχέτισης διήρκησε για δύο χρονικά διαστήματα (έξι μήνες) αλλά δεν εμπίπτει στους σκοπούς της παρούσας διπλωματικής εργασίας.

Βιβλιογραφία

[1] Newman, M.. (2011). Communities, Modules and Large-Scale Structure in Networks. Nature Physics. 8. 25-31. 10.1038/nphys2162.

[2] Fortunato, Santo. (2009). Community Detection in Graphs. Physics Reports. 486. 10.1016/j.physrep.2009.11.002.

[3] Mancoridis, Spiros & Mitchell, Brian & Rorres, Chris & Chen, Yih-Farn & Gansner, Emden. (1998). Using Automatic Clustering to Produce High-Level System Organizations of Source Code. Program Comprehension, Workshop Proceedings. 45 - 52. 10.1109/WPC.1998.693283.

[4] Andrews, G. E., 1976, The theory of partitions (Addison-Wesley, Boston, USA).

[5] Μωυσιάδης Πολυχρόνης, (2002), Συνδυαστική Απαρίθμηση (ZHTH, Θεσσαλονίκη, Ελλάδα)

[6] Krishnamurthy, Balachander & Wang, Jia. (2001). On Network-Aware Clustering of Web Clients. ACM SIGCOMM Computer Communication Review. 30. 10.1145/347059.347412.

[7] Reddy, K. P., M. Kitsuregawa, P. Sreekanth, and S. S. Rao, 2002, in DNIS '02: Proceedings of the Second International Workshop on Databases in Networked Information Systems (Springer-Verlag, London, UK), pp. 188-200.

[8] Wu, A. Y., M. Garland, and J. Han, 2004, in KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (ACM Press,New York, NY, USA), pp. 719-724.

[9] Freeman, Linton. (1977). A Set of Measures of Centrality Based on Betweenness. Sociometry. 40. 35-41. 10.2307/3033543.

[10] Granovetter, Mark. (1973). The Strength of Weak Ties. The American Journal of Sociology. 78. 1360-1380. 10.1086/225469.

[11] Guimerà, Roger & Amaral, Luís. (2005). Functional Cartography of Complex Metabolic Networks. Nature. 23. 22-231.

[12] Simon, Herbert. (1962). The Architecture of Complexity. Proceedings of the American Philosophical Society. 106. 467-482. 10.1007/978-3-642-27922-5_23.

[13] Newman, Mark & Girvan, Michelle. (2004). Finding and Evaluating Community Structure in Networks. Physical review. E, Statistical, nonlinear, and soft matter physics. 69. 026113. 10.1103/PhysRevE.69.026113.

[14] Newman, M. (2004). Analysis of Weighted Networks. Physical review. E, Statistical, nonlinear, and soft matter physics. 70. 056131. 10.1103/PhysRevE.70.056131.

[15] Leicht, EA & Newman, M. (2008). Community Structure in Directed Networks. Physical review letters. 100. 118703. 10.1103/PhysRevLett.100.118703.

[16] Gomez, Sergio & Jensen, Pablo & Arenas, Alex. (2009). Analysis of community structure in networks of correlated data. Physical review. E, Statistical, nonlinear, and soft matter physics. 80. 016114. 10.1103/PhysRevE.80.016114.

[17] Sarzynska, Marta & Leicht, Elizabeth & Chowell, Gerardo & Porter, Mason. (2014). Null Models for Community Detection in Spatially-Embedded, Temporal Networks. Journal of Complex Networks. 4. 10.1093/comnet/cnv027.

[18] Liu, Xin & Murata, Tsuyoshi & Wakita, Ken. (2012). Extending modularity by capturing the similarity attraction feature in the null model.

[19] Chang, Yu-Teng & Leahy, Richard & Pantazis, Dimitrios. (2012). Modularity-based graph partitioning using conditional expected models. Physical review. E, Statistical, nonlinear, and soft matter physics. 85. 016109.

10.1103/PhysRevE.85.016109.

[20] https://en.wikipedia.org/wiki/Spatial_network (date of access 1/7/2019).

[21] Expert, Paul & Evans, Tim & Blondel, Vincent & Lambiotte, Renaud. (2011). Uncovering Space-Independent Communities in Spatial Networks. Proceedings of the National Academy of Sciences of the United States of America. 108. 7663-8. 10.1073/pnas.1018962108.

[22] Simini, Filippo & Gonzalez, Marta C. & Maritan, Amos & Barabasi, Albert-Laszlo. (2012). A Universal Model for Mobility and Migration Patterns. Nature. 484. 96-100. 10.1038/nature10856.

[23] Macmahon, Mel & Garlaschelli, Diego. (2013). Community Detection for Correlation Matrices. Physical Review X. 5. 10.1103/PhysRevX.5.021006.

[24] Brandes, Ulrik & Delling, Daniel & Gaertler, Marco & Görke, Robert & Hoefer, Martin & Nikoloski, Zoran & Wagner, Dorothea. (2006). On Modularity - NP-Completeness and Beyond., URL http://digbib.ubka.uni-karlsruhe.de/volltexte/documents/3255. (date of access 1/7/2019)

[25] Blondel, Vincent & Guillaume, Jean-Loup & Lambiotte, Renaud & Lefebvre, Etienne. (2008). Fast Unfolding of Communities in Large Networks. Journal of Statistical Mechanics Theory and Experiment. 2008. 10.1088/1742-5468/2008/10/P10008.

[26] https://en.wikipedia.org/wiki/Greedy_algorithm (date of access 1/7/2019).

[27] Lancichinetti, Andrea & Fortunato, Santo. (2009). Community Detection Algorithms: A Comparative Analysis. Physical review. E, Statistical, nonlinear, and soft matter physics. 80. 056117. 10.1103/PhysRevE.80.056117.

[28] Fortunato, Santo & Barthelemy, Marc. (2007). Resolution limit in community detection. Proceedings of the National Academy of Sciences of the United States of America. 104. 36-41. 10.1073/pnas.0605965104.

[29] Sporns, Olaf & Betzel, Richard. (2015). Modular Brain Networks. Annual review of psychology. 67. 10.1146/annurev-psych-122414-033634.

[30] Reichardt, Jörg & Bornholdt, Stefan. (2006). Bornholdt, S.: Statistical Mechanics of Community Detection. Physics Review E 74(1), 016110.

[31] He, Zengyou & Liang, Hao & Chen, Zheng & Zhao, Can. (2018). Detecting Statistically Significant Communities. arXiv:1806.05602

[32] Karrer, Brian & Levina, Elizaveta & Newman, M. (2008). Robustness of community structure networks. Physical review. E, Statistical, nonlinear, and soft matter physics. 77. 046119. 10.1103/PhysRevE.77.046119.

[33] Hu, Yanqing & Nie, Yuchao & Yang, Hua & Cheng, Jie & Fan, Ying & Di, Zengru. (2010). Measuring Significance of Community Structure in Complex Networks. Physical review. E, Statistical, nonlinear, and soft matter physics. 82. 066106. 10.1103/PhysRevE.82.066106.

[34] Carissimo, Annamaria & Cutillo, Luisa & Feis, Italia. (2018). Validation of community robustness. Computational Statistics & Data Analysis. 120. 10.1016/j.csda.2017.10.006.

[35] Spirin, Victor & Mirny, Leonid. (2003). Spirin V, Mirny LA.. Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci USA 100: 12123-12128. Proceedings of the National Academy of Sciences of the United States of America. 100. 12123-8. 10.1073/pnas.2032324100.

[36] Kojaku, Sadamori & Masuda, Naoki. (2017). A generalised significance test for individual communities in networks. Scientific Reports. 8. 10.1038/s41598-018-25560-z.

[37] Newman, M.E.J.. (2003). Mixing patterns in networks. Physical review. E, Statistical, nonlinear, and soft matter physics. 67. 026126. 10.1103/PhysRevE.67.026126.

[38] Milo, Ron & Kashtan, N & Itzkovitz, Shalev & Newman, M. & Alon, Uri. (2004). On the uniform generation of random graphs with prescribed degree sequences. Tech rep. 21.

[39] Maslov, Sergei & Sneppen, Kim & Zaliznyak, Alexei. (2004). Detection of topological patterns in complex networks: Correlation profile of the internet. Physica A: Statistical Mechanics and its Applications. 333. 529-540. 10.1016/j.physa.2003.06.002.

[40] https://en.wikipedia.org/wiki/Degree-preserving_randomization (date of access 1/7/2019).

[41] Trusina, Ala & Maslov, Sergei & Minnhagen, Petter & Sneppen, Kim. (2004). Hierarchy Measures in Complex Networks. Physical review letters. 92. 178702. 10.1103/PhysRevLett.92.178702.

[42] Fagiolo, Giorgio & Squartini, Tiziano & Garlaschelli, Diego. (2012). Null Models of Economic Networks: The Case of the World Trade Web. Journal of Economic Interaction and Coordination. 8. 75-107. 10.1007/s11403-012-0104-7.

[43] Ansmann, Gerrit & Lehnertz, Klaus. (2011). Constrained randomization of weighted networks. Physical review. E, Statistical, nonlinear, and soft matter physics. 84. 026103. 10.1103/PhysRevE.84.026103.

[44] Genio, Charo & Kim, Hyunju & Toroczkai, Zoltan & Bassler, Kevin. (2010). Efficient and Exact Sampling of Simple Graphs with Given Arbitrary Degree Sequence. PloS one. 5. e10012. 10.1371/journal.pone.0010012.

[45] Chorozoglou, Dimitrios & Kugiumtzis, Dimitris. (2018). Testing the randomness of correlation networks from multivariate time series. *Journal of Complex Networks*, Volume 7, Issue 2, April 2019, Pages 190–209.

[46] Zalesky, Andrew & Fornito, Alex & Bullmore, Edward. (2012). On the use of correlation as a measure of network connectivity. NeuroImage. 60. 2096-106. 10.1016/j.neuroimage.2012.02.001.

[47] Kugiumtzis, Dimitris. (2002). Surrogate Data Test on Time Series. 10.1007/978-1-4615-0931-8_13.

[48] Theiler, James & Prichard, Dean. (1996). Using "Surrogate Surrogate Data" to Calibrate the Actual Rate of False Positives in Tests for Nonlinearity in Time Series.

[49] Horvath, Steve. (2011). Weighted Network Analysis. 10.1007/978-1-4419-8819-5_1.

[50] Hirschberger, Markus & Qi, Yue & Steuer, Ralph. (2007). Randomly Generating PortfolioSelection Covariance Matrices with Specified Distributional Characteristics. European Journal of Operational Research. 177. 1610-1625. 10.1016/j.ejor.2005.10.014.

[51] Fornito, A. & Zalesky, A. & Bullmore, Edward. (2016). Fundamentals of Brain Network Analysis.

[52] Hosseini, Hadi & Kesler, Shelli. (2013). Influence of Choice of Null Network on Small-World Parameters of Structural Correlation Networks. PloS one. 8. e67354. 10.1371/journal.pone.0067354.

[53] D. Chorozoglou, D. Kugiumtzis, “Testing the randomness of causality networks from multivariate time series”, 2014 International Symposium on Nonlinear Theory and its Applications (NOLTA 2014), Luzern, 229-232, 2014

[54] Kugiumtzis, Dimitris. (2002). Statically Transformed Autoregressive Process and Surrogate Data Test for Nonlinearity. Physical review. E, Statistical, nonlinear, and soft matter physics. 66. 025201. 10.1103/PhysRevE.66.025201.

[55] Schreiber T. & Schmitz A. (1996) Improved surrogate data for nonlinearity tests. Phys. Rev. Lett., 77, 635-638.

[56] https://www.mathworks.com/matlabcentral/fileexchange/45867-community-detection-toolbox?s_tid=prof_contriblnk (date of access 1/7/2019).

[57] http://users.auth.gr/dkugiu/ (date of access 1/7/2019).

ΠΑΡΑΡΤΗΜΑ

Παραθέτονται οι αντίστοιχοι πίνακες της παραγράφου 5.1.4 για στάθμη σημαντικότητας $\alpha = 0.01$

Πίνακας 6: Πιθανότητες απόρριψης της μηδενικής υπόθεσης T=200, N=50, πλήθος κοινοτήτων = 3 (25-15-10), a=0.01

A	in block S	out block S	H_0	maslov	Iaaft bin	Stap bin	ansmann	Iaaft wei	Stap wei
U[0.01 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.2]	Ψευδής	1	1	1	1	1	1
U[0.01 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.4]	Ψευδής	0.94	0.91	0.94	1	1	1
U[0.01 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.6	Αληθής	0.3	0.19	0.18	0.99	1	1
U[0.01 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.9]	Αληθής	0.16	0.09	0.10	0.98	0.99	1
U[0.01 , 0.9]	U[0.01 , 0.4]	U[0.01 , 0.4]	Αληθής	0.24	0.11	0.09	0.89	0.97	0.97
U[0.01 , 0.9]	U[0.01 , 0.3]	U[0.01 , 0.3]	Αληθής	0.23	0.08	0.16	0.89	0.95	0.96
U[0.01 , 0.9]	U[0.01 , 0.25]	U[0.01 , 0.25]	Αληθής	0.14	0.09	0.11	0.87	0.95	0.96
U[0.5 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.2]	Ψευδής	1	1	1	1	1	1
U[0.5 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.4]	Ψευδής	0.97	0.94	0.90	1	1	1
U[0.5 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.6	Αληθής	0.18	0.08	0.11	0.94	0.98	0.98
U[0.5 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.9]	Αληθής	0.32	0.13	0.13	0.82	0.88	0.91
U[0.5 , 0.9]	U[0.01 , 0.4]	U[0.01 , 0.4]	Αληθής	0.28	0.12	0.12	0.76	0.86	0.88
U[0.5 , 0.9]	U[0.01 , 0.3]	U[0.01 , 0.3]	Αληθής	0.08	0.03	0.03	0.53	0.72	0.73
U[0.5 , 0.9]	U[0.01 , 0.25]	U[0.01 , 0.25]	Αληθής	0.07	0.03	0.04	0.60	0.70	0.75

Πίνακας 7: Πιθανότητες απόρριψης της μηδενικής υπόθεσης T=200, N=50, πλήθος κοινοτήτων = 2, (25-25), a=0.01

A	in block S	out block S	H_0	maslov	Iaaft bin	Stap bin	ansmann	Iaaft wei	Stap wei
U[0.01 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.2]	Ψευδής	1	1	1	1	1	1
U[0.01 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.4]	Ψευδής	0.99	1	0.99	1	1	1
U[0.01 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.6	Αληθής	0.53	0.46	0.39	1	1	1
U[0.01 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.9]	Αληθής	0.21	0.07	0.10	0.97	0.99	1
U[0.01 , 0.9]	U[0.01 , 0.4]	U[0.01 , 0.4]	Αληθής	0.24	0.09	0.13	0.98	0.99	0.99
U[0.5 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.2]	Ψευδής	1	1	1	1	1	1
U[0.5 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.4]	Ψευδής	0.88	0.84	0.82	1	1	1
U[0.5 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.6	Αληθής	0.44	0.21	0.29	0.92	0.96	0.97
U[0.5 , 0.9]	U[0.01 , 0.9]	U[0.01 , 0.9]	Αληθής	0.39	0.10	0.11	0.83	0.89	0.89
U[0.5 , 0.9]	U[0.01 , 0.4]	U[0.01 , 0.4]	Αληθής	0.22	0.09	0.12	0.75	0.93	0.91

Πίνακας 8: Πιθανότητες απόρριψης της μηδενικής υπόθεσης T=200, N=50, a=0.01, σταθερές τιμές για τα στοιχεία του πίνακα συνδιασπορών.

A	S	H_0	maslov	Iaaft bin	Stap bin	ansmann	Iaaft wei	Stap wei
U[0.01 , 0.9]	0.2	Αληθής	0.01	0.01	0	1	1	1
U[0.01 , 0.9]	0.25	Αληθής	0.08	0.01	0.02	1	1	1
U[0.01 , 0.9]	0.3	Αληθής	0.11	0.05	0.05	0.90	0.97	0.98
U[0.5 , 0.9]	0.2	Αληθής	0.10	0	0.02	1	1	1
U[0.5 , 0.9]	0.25	Αληθής	0.09	0.01	0.02	0.95	0.97	0.98
U[0.5 , 0.9]	0.3	Αληθής	0.07	0.02	0.03	0.64	0.72	0.75