ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ Μ/Υ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΣΧΟΛΗ ΝΑΥΤΙΛΙΑΣ ΚΑΙ ΒΙΟΜΗΧΑΝΙΑΣ
ΤΜΗΜΑΤΟΣ ΒΙΟΜΗΧΑΝΙΚΗΣ ΔΙΟΙΚΗΣΗΣ & ΤΕΧΝΟΛΟΓΙΑΣ
ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΤΕΧΝΟ-ΟΙΚΟΝΟΜΙΚΑ ΣΥΣΤΗΜΑΤΑ»

Ανάλυση Επιχειρηματικών Δικτύων με Αλγορίθμους Ανακάλυψης Κοινοτήτων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΝΙΚΟΛΕΤΑ- ΑΦΡΟΔΙΤΗ ΠΑΝΑΓΟΥ

ΕΠΙΒΛΕΠΩΝ: Συμεών Παπαβασιλείου
Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2020

ΠΕΡΙΛΗΨΗ

Η παρούσα Μεταπτυχιακή Διπλωματική Εργασία επικεντρώνεται στην ανάπτυξη μιας μεθοδολογίας ανάλυσης δεδομένων επιχειρήσεων για την βελτιστοποίηση της διαδικασίας λήψης επιχειρηματικών αποφάσεων. Ειδικότερα, το πλαίσιο που προτείνεται αξιοποιεί πληροφορία που αφορά τόσο τις αλληλεπιδράσεις της ίδιας της επιχειρηματικής οντότητας με άλλες, όσο και τις συσχετίσεις που αναπτύσσονται στο ευρύτερο επιχειρηματικό περιβάλλον αυτής, σε αντίθεση με τις παραδοσιακές μεθόδους που αντιμετωπίζουν κάθε οντότητα μεμονωμένα. Η προτεινόμενη μεθοδολογία εφαρμόζεται στο παγκόσμιο δίκτυο νεοφυών επιχειρήσεων, όπου δεν υπάρχει επαρκής διαχρονική πληροφορία για ανάλυση τάσεων και μοντέλα προβλέψεων, επομένως η απόφαση πρέπει να υποστηριχθεί με διαφορετικά μοντέλα και τεχνικές.

Για την ανάπτυξη της μεθοδολογίας χρησιμοποιούνται τεχνικές Ανάλυσης Κοινωνικών Δικτύων και συγκεκριμένα δυο μέθοδοι. Αρχικά επιλέγεται ο αλγόριθμος ανίχνευσης κοινοτήτων DBSCAN (Density- Based Spatial Clustering of Applications with Noise), τροποποιημένος ως προς τη μετρική της απόστασης που χρησιμοποιεί. Ειδικότερα, προτείνεται η ομαδοποίηση με βάση την υπερβολική απόσταση μεταξύ των οντοτήτων και για αυτό το λόγο γίνεται ενσωμάτωση του δικτύου των επιχειρήσεων στον υπερβολικό χώρο. Στη συνέχεια εφαρμόζεται ο αλγόριθμος Greedy Modularity Communities, ο οποίος ανιχνεύει κοινότητες μέσα στο δίκτυο με κριτήριο τη μεγιστοποίησης της αρθρωτότητας.

Οι κοινότητες που προέκυψαν από τους δυο αλγορίθμους μελετήθηκαν, τόσο ως προς τα χαρακτηριστικά των κόμβων, όσο και ως προς τη δομή της κοινότητας και του δικτύου στο σύνολο. Έτσι, μέσα από την εφαρμογή αυτή, επιχειρείται να εξαχθούν συμπεράσματα για το αν οι μέθοδοι ανάλυσης κοινωνικών δικτύων μπορούν να συνεισφέρουν στην ανακάλυψη μοτίβων ή «κρυφών» ιδιοτήτων» που υπάρχουν μέσα σε ένα δίκτυο επιχειρήσεων και δυνητικά μπορούν να καθορίσουν τις αποφάσεις.

Λέξεις Κλειδιά: Κοινωνικά Δίκτυα, Ανίχνευση Κοινοτήτων, Δίκτυο Επιχειρήσεων, Υπερβολική Γεωμετρία, Ενσωμάτωση Δικτύου

ABSTRACT

The present Master Thesis is focused on the development of a methodology for analyzing business data, with a view to optimize the business decision-making process. Our approach considers the business entity as part of a network, in which every interaction is taken into account, regardless of whether or not it affects the business directly. More precisely, the proposed framework is based on information both within and between the business environments, and is implemented at the «Worldwide Startup Network». The last one was chosen for testing the methodology, as representative of those cases where there are limitations for trend analysis and forecasting models, and therefore the decision must be supported by different techniques.

To this end, two Social Network Analysis techniques are used for detecting communities in the World Wide Startup Network, a) the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm and b) a Modularity Maximization Method. The former is modified in order to adopt hyperbolic geometry and for this purpose the network is embedded in Hyperbolic Space. From the Modularity Maximization Methods, Greedy Modularity Communities algorithm was chosen as the most efficient concerning the size of the network. The results of the two algorithms were evaluated through analyzing the companies' attributes and communities' structure.

Along these lines, the present thesis is trying to explore weather Social Network Analysis techniques can reveal patterns or hidden properties in the business network that could potentially lead the decision making process.

Keywords: Social Networks, Community Detection, Business Network, Hyperbolic Geometry, Network Embedding

Πρόλογος

Η εργασία αυτή εκπονήθηκε κατά το ακαδημαϊκό έτος 2019-2020 στο πλαίσιο του Προγράμματος Μεταπτυχιακών Σπουδών «Τεχνοοικονομικά Συστήματα».

Πρωτίστως θα ήθελα να ευχαριστήσω τον Καθηγητή της σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών κύριο Συμεών Παπαβασιλείου για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα τόσο ενδιαφέρον αντικείμενο.

Ιδιαίτερες ευχαριστίες οφείλω στον υποψήφιο διδάκτορα Κωνσταντίνο Τσιτσεκλή και στον Αναπληρωτή Καθηγητή του Τμήματος Πληροφορικής του Ιονίου Πανεπιστημίου, κύριο Βασίλειο Καρυώτη, για το χρόνο που αφιέρωσαν στην επίβλεψη της εργασίας αυτής, καθώς και στην καθοδήγηση που μου παρείχαν σε όλη τη διάρκεια της εκπόνησης.

Η μεταπτυχιακή αυτή εργασία αφιερώνεται στην οικογένεια και τους φίλους μου, ως δείγμα ευγνωμοσύνης για τη στήριξή τους.

Περιεχόμενα

Κατάλογος Εικόνων

Κατάλογος Πινάκων

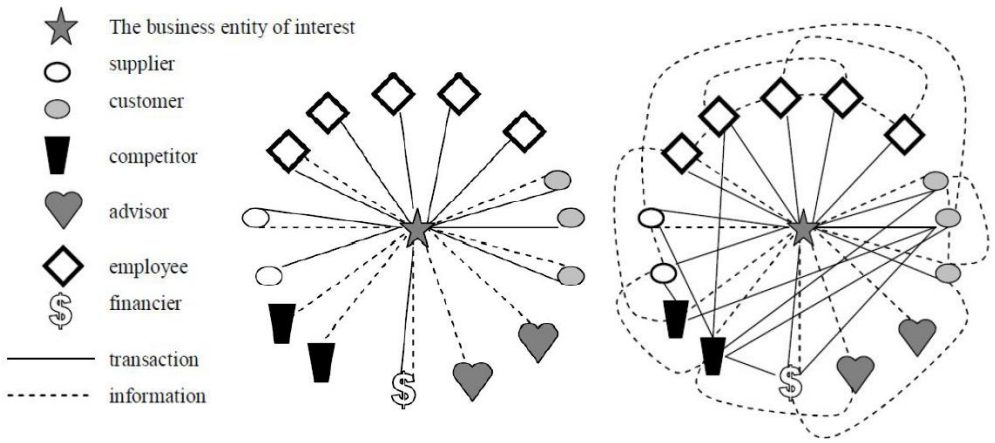
Κατάλογος Γραφημάτων

1

Εισαγωγή

1.1 Ανάλυση Σύνθετων Δικτύων και Δίκτυα Επιχειρήσεων

Στην επιστήμη της Διοίκησης Επιχειρήσεων, ως δίκτυο επιχειρήσεων ορίζεται ένα σύστημα οντοτήτων που αλληλεπιδρούν. Στο μεγαλύτερο μέρος της βιβλιογραφίας, η ανάλυση των δικτύων επιχειρήσεων επικεντρώνεται σε εγωκεντρικά δίκτυα (ego-networks), όπου μελετάται η σύνδεση μιας επιχειρηματικής οντότητας (ego) με διάφορους παράγοντες (alters), με τους οποίους η οντότητα έχει άμεσες σχέσεις-συναλλαγές. Αυτοί μπορεί να είναι πελάτες, προμηθευτές, δανειστές, ανταγωνιστές κ.λπ. [1]. Η σύγχρονη οικονομική θεωρία υιοθετεί αυτό που στην επιστήμη της θεωρίας γραφημάτων ονομάζεται πλήρες δίκτυο, το οποίο εξετάζει τις άμεσες και έμμεσες συσχετίσεις μεταξύ όλων των οντοτήτων. Στην εικόνα 1.1 φαίνεται η διαφορά ανάμεσα στην κλασική και την πλήρη γραφοθεωρητική προσέγγιση του δικτύου επιχειρήσεων.



Εικόνα 1.1 *Εγωκεντρικό (ego) και πλήρες (complete) δίκτυο επιχειρήσεων* **[1]**

Η πολυπλοκότητα και το πλήθος των σχέσεων ανάμεσα στις οντότητες των σύγχρονων δικτύων επιχειρήσεων τα κατατάσσει στην κατηγορία των Σύνθετων Δίκτυων (Complex Networks) και η μελέτη τους αποτελεί αντικείμενο της «Ανάλυσης Κοινωνικών Δικτύων». Τέτοια δίκτυα παρουσιάζουν ενδιαφέρουσες τοπολογικές ιδιότητες, όπως για παράδειγμα η τάση που έχουν οι κόμβοι να σχηματίζουν κοινότητες, δηλαδή σύνολα κόμβων που συνδέονται μεταξύ τους με περισσότερες συνδέσεις από ότι με κόμβους άλλων ομάδων. Με βάση την παραδοχή ότι η δομή του δικτύου επιχειρήσεων επηρεάζει τη δύναμη και τη θέση κάθε μέλους του στην αγορά, η εφαρμογή τεχνικών Ανάλυσης Κοινωνικών Δικτύων μπορεί να δώσει σημαντικές πληροφορίες για τη συνοχή του δικτύου και τη θέση/σημαντικότητα μεμονωμένων κόμβων ή ομάδων μέσα σε αυτό.

1.2 Αντικείμενο Διπλωματικής Εργασίας

Τα τελευταία χρόνια καταγράφεται μεγάλο ενδιαφέρον γύρω από τις νεοφυείς επιχειρήσεις (startups) παγκοσμίως. Νεοφυής επιχείρηση είναι ένας οργανισμός που συνδέεται με υψηλή ανάπτυξη, έχει συνήθως τεχνολογικό προσανατολισμό, καινοτόμο επιχειρηματικό μοντέλο (business model) και στόχος της είναι να εδραιώσει μια νέα αγορά ή να επεκτείνει μια ήδη υπάρχουσα [2]. Οι επιχειρήσεις αυτού του είδους παρουσιάζουν χαμηλό κόστος υλοποίησης αλλά πολύ υψηλό ρίσκο και αντίστοιχα πολύ υψηλή απόδοση στην περίπτωση επιτυχίας. Όλα τα παραπάνω, σε συνδυασμό με το γεγονός ότι οι επιτυχημένες νεοφυείς επιχειρήσεις έχουν πολύ

μεγαλύτερη δυνατότητα επεκτασιμότητας (λόγω των χαμηλών απαιτήσεων σε κεφάλαιο), τις καθιστούν εξαιρετικά ελκυστικές για τους επενδυτές, σε σύγκριση με τις τυπικές επιχειρήσεις.

Ωστόσο, λόγω της αβεβαιότητας που ενέχουν ως επιχειρηματική δραστηριότητα, είναι δύσκολο να εφαρμοσθούν τα παραδοσιακά μοντέλα αξιολόγησης επενδύσεων, τα οποία βασίζονται στις τάσεις που εξάγονται από ιστορικά δεδομένα (όπως για παράδειγμα πωλήσεις, κατανάλωση, μέγεθος αγοράς, παραγωγική ικανότητα, κ.λπ.). Αυτό οδηγεί τους επενδυτές να αξιολογούν τις επιχειρήσεις αυτές βασιζόμενοι στις προοπτικές της κεντρικής επιχειρηματικής ιδέας, στις προσδοκίες που έχουν από τον ιδιοκτήτη της επιχείρησης και γενικότερα σε υποκειμενικά κριτήρια που ενέχουν μεγάλη αβεβαιότητα και προκατάληψη [3].

Σύμφωνα με τα παραπάνω, λοιπόν, η ανάλυση του δικτύου των νεοφυών επιχειρήσεων με τεχνικές Ανάλυσης Κοινωνικών Δικτύων, λαμβάνοντας υπόψη όλη τη διαθέσιμη πληροφορία για κάθε νεοφυή επιχείρηση, αλλά και όλες τις ιδιότητες των συνδέσεων μεταξύ τους, μπορεί να προσδιορίσει μοτίβα (patterns), μέτρα δύναμης και επιρροής, αλλά και κοινότητες μέσα στο δίκτυο, και να παρέχουν πιο αντικειμενικά κριτήρια για την αξιολόγηση επενδυτικών σχεδίων.

1.2.1 Συνεισφορά

Στην παρούσα εργασία προτείνεται ένα πλαίσιο ανάλυσης δικτύων επιχειρήσεων που λαμβάνει υπόψη τον τεράστιο όγκο της διαθέσιμης πληροφορίας που παρέχουν διαδικτυακές πλατφόρμες με σκοπό τη βελτιστοποίηση της λήψης επιχειρηματικών αποφάσεων. Για τη διαχείριση του όγκου της πληροφορίας που συνοδεύει τις νεοφυείς επιχειρήσεις, αλλά και για την κατασκευή του δικτύου, έγινε χρήση μη σχεσιακής βάσης δεδομένων. Για τη διασύνδεση των επιχειρήσεων, πέρα από τη λογική των συναλλαγών (transactions) που καθιστούν δύο επιχειρήσεις συνδεόμενες, εισάγονται και δύο ακόμη παράμετροι (features) που είναι η ροή α) τεχνογνωσίας και β) επενδυτικού ενδιαφέροντος μέσα στο δίκτυο.

Μετά την κατασκευή του δικτύου προτείνεται η ενσωμάτωσή του στον υπερβολικό χώρο, με στόχο αφενός τη μείωση διαστάσεων των δεδομένων, και αφετέρου μια πιο κατανοητή αποτύπωση της απόστασης μεταξύ δυο κόμβων. Έτσι οι κόμβοι αποκτούν

συντεταγμένες με τρόπο ώστε η μεταξύ τους απόσταση να προσεγγίζει το συντομότερο μονοπάτι ανάμεσά τους.

Στη συνέχεια, προτείνεται η εφαρμογή δύο αλγορίθμων για την ανίχνευση κοινοτήτων μέσα στο δίκτυο. Έχοντας ένα δίκτυο με γεωμετρική πληροφορία (συντεταγμένες), εφαρμόζεται ο αλγόριθμος ομαδοποίησης DBSCAN, ο οποίος συνιστά μη επιβλεπόμενη μέθοδο ομαδοποίησης (clustering) και βασίζεται στις αποστάσεις μεταξύ των οντοτήτων ενός χωρικού συστήματος. Ο αλγόριθμος αυτός τροποποιήθηκε μερικώς, έτσι ώστε να χρησιμοποιεί μια συνάρτηση υπολογισμού υπερβολικής απόστασης μεταξύ των κόμβων του δικτύου. Ακολούθως εφαρμόσθηκε η μέθοδος μεγιστοποίησης της αρθρωτότητας (Modularity Maximization) ώστε να βρεθεί ο βέλτιστος διαχωρισμός κοινοτήτων μέσα στο δίκτυο.

1.3 Οργάνωση Κειμένου

Η παρούσα εργασία οργανώνεται σε 9 κεφάλαια. Το παρόν κεφάλαιο ορίζει το πρόβλημα και θέτει το στόχο της εργασίας, ενώ τα επόμενα πέντε κεφάλαια (έως και το 6^{ο}) αναλύουν βασικές έννοιες και αποσαφηνίζουν τη μεθοδολογική προσέγγιση της επίλυσης. Στο 7^{ο} κεφάλαιο παρατίθενται οι μέθοδοι ομαδοποίησης που εφαρμόσθηκαν, στο 8^{ο} κεφάλαιο παρουσιάζονται τα αποτελέσματα και στο 9^{ο} κεφάλαιο παρατίθενται τα συμπεράσματα από την εφαρμογή των δυο μεθόδων. Ειδικότερα:

Στο **Κεφάλαιο 2** καταγράφεται το απαραίτητο θεωρητικό υπόβαθρο και η ορολογία για την κατανόηση της μεθοδολογίας που ακολουθείται. Γίνεται αναφορά σε βασικές έννοιες της θεωρίας γραφημάτων.

Στο **Κεφάλαιο 3** αναφέρονται ιδιότητες και χαρακτηριστικά των σύνθετων δικτύων που κρίνονται απαραίτητα για την ανάλυσή των δομικών στοιχείων τους.

Ακολούθως, στο **Κεφάλαιο 4** καταγράφεται ο τρόπος με τον οποίο αξιοποιήθηκαν διαδικτυακά δεδομένα για να κατασκευασθεί το παγκόσμιο δίκτυο νεοφυών επιχειρήσεων.

Το **Κεφάλαιο 5** αναφέρεται στη διαδικασία ανακάλυψης κοινοτήτων σε σύνθετα δίκτυα, ενώ αναλύεται και ο τρόπος λειτουργίας των αλγορίθμου που επιλέχθηκαν για τη συγκεκριμένη εφαρμογή.

Στη συνέχεια, στο **Κεφάλαιο 6** καταγράφονται βασικά στοιχεία της υπερβολικής γεωμετρίας και αναφέρονται οι ιδιότητες που καθιστούν τη χρήση της σημαντική για τη συγκεκριμένη εφαρμογή.

Στο **Κεφάλαιο 7** αναλύονται οι προτεινόμενες μέθοδοι ανακάλυψης κοινοτήτων σε και επεξηγείται ο τρόπος λειτουργίας τους.

Το **Κεφάλαιο 8** επικεντρώνεται στα αποτελέσματα της εφαρμογής των μεθόδων στο παγκόσμιο δίκτυο νεοφυών επιχειρήσεων.

Τέλος, στο Κ**εφάλαιο 9** καταγράφονται τα συμπεράσματα και γίνονται προτάσεις για περαιτέρω έρευνα.

2

Εισαγωγικές Έννοιες Θεωρίας Γραφημάτων

Για τη μελέτη σύνθετων δικτύων όπως είναι τα βιολογικά ή εν προκειμένω τα κοινωνικά, και πολλά ακόμη, είναι απαραίτητη μια μαθηματική αναπαράσταση, η οποία να αποδίδει τις σχέσεις μεταξύ των οντοτήτων με έναν εύληπτο και σαφή τρόπο. Η λύση στο πρόβλημα δίνεται από τον τομέα των διακριτών μαθηματικών και τη Θεωρία Γραφημάτων. Στη συνέχεια του κεφαλαίου παρατίθενται βασικές έννοιες του εργαλείου αυτού, οι οποίες χρησιμοποιούνται στο πλαίσιο της εργασίας.

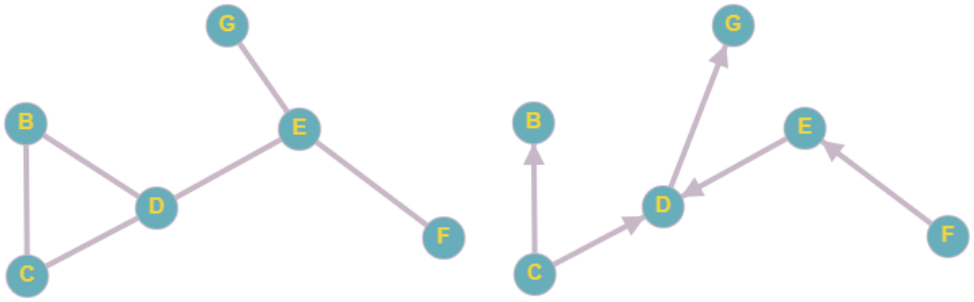
2.1 Βασική Ορολογία

Το γράφημα ή γράφος είναι ένας τρόπος μοντελοποίησης ενός δικτύου. Ορίζεται ως ένα ζεύγος $G = (V, E)$, όπου V είναι ένα πεπερασμένο- μη κενό σύνολο κόμβων (vertices) και E είναι ένα πεπερασμένο σύνολο ζευγών με στοιχεία του συνόλου V, τα οποία ονομάζονται ακμές (edges).

Υπογράφος (subgraph) του $G = (V, E)$ είναι ένας γράφος $G_0 = (V_0, E_0)$ για τα στοιχεία του οποίου ισχύει $V_0 \subseteq V$, $E_0 \subseteq E$ και το E_0 ορίζεται στο V_0.

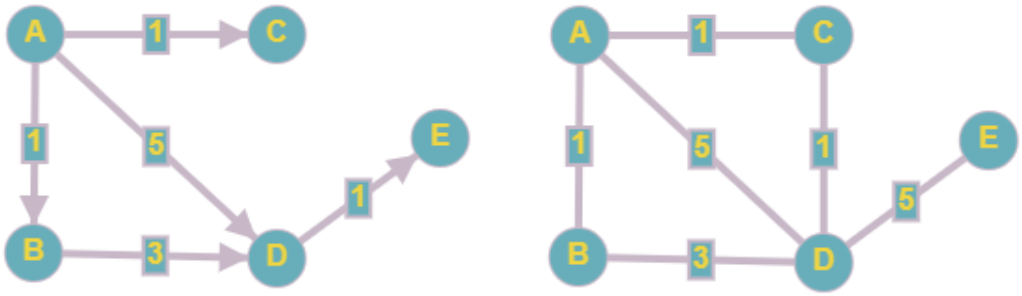
Τα ζεύγη στοιχείων του συνόλου V μπορεί να είναι διατεταγμένα ή μη διατεταγμένα. Στην περίπτωση που είναι διατεταγμένα, οι ακμές έχουν προσανατολισμό και το γράφημα ονομάζεται κατευθυνόμενο (directed). Σε αντίθετη περίπτωση, όταν τα ζεύγη κόμβων είναι μη διατεταγμένα, οι ακμές δεν έχουν προσανατολισμό και το

γράφημα ονομάζεται μη κατευθυνόμενο (undirected). Στην εικόνα 2.1 , φαίνονται οι δύο τύποι γραφημάτων.



Εικόνα 2.1 *Απεικόνιση ενός μη κατευθυνόμενου γράφου (αριστερά) και ενός κατευθυνόμενου υπογράφου (δεξιά) .*

Ένας γράφος, εκτός από κατευθυνόμενος και μη κατευθυνόμενος, μπορεί να είναι και γράφος με βάρη ή χωρίς. Γράφος με βάρη (weighted graph) είναι εκείνος του οποίου οι ακμές αντιπροσωπεύουν μια μετρήσιμη ποσότητα (π.χ. απόσταση, κόστος, χωρητικότητα, κ.α.). Η ποσότητα αυτή προσδιορίζεται από έναν αριθμό ο οποίος επισυνάπτεται στις ακμές. Στην εικόνα 2.2 φαίνονται κατά σειρά ένας κατευθυνόμενος και ένας μη κατευθυνόμενος γράφος με βάρη.



Εικόνα 2.2 *Απεικόνιση κατευθυνόμενου (αριστερά) και μη κατευθυνόμενου (δεξιά) γραφήματος με βάρη*

Ένας μαθηματικός τρόπος για να αναπαρασταθεί ένα γράφημα είναι ο πίνακας γειτνίασης (adjacency matrix). Πρόκειται για τετραγωνικό πίνακα Α, διαστάσεων n×n, όπου n ο αριθμός των κόμβων. Κάθε στοιχείο a_{ij} του Α αντιπροσωπεύει την ύπαρξη ή όχι ακμής μεταξύ δύο κόμβων i και j. Αν οι κόμβοι i,j συνδέονται με ακμή, τότε θεωρούνται γείτονες και a_{ij} =1 (στην περίπτωση που ο γράφος είναι χωρίς βάρη). Αν αντιθέτως πρόκειται για γράφο με βάρη, τότε για δυο γειτονικούς κόμβους το στοιχείο a_{ij} παίρνει την τιμή του βάρους της αντίστοιχης ακμής. Σε κάθε περίπτωση,

αν δεν υπάρχει ακμή που να συνδέει δυο κόμβους, η αντίστοιχη τιμή στον πίνακα γειτνίασης είναι μηδενική.

2.2 Περίπατοι- Μονοπάτια

Περίπατος (walk) ονομάζεται μια ακολουθία κόμβων $W = (v_0, v_1, \dots, v_k)$ του γραφήματος $G = (V, E)$ εάν $v_{i-1}v_i \in E$ για κάθε i=1,2,…,k. Το μήκος του περιπάτου ισούται με το πλήθος των ακμών από τις οποίες αποτελείται.

Διαδρομή ή μονοπάτι (path) ονομάζεται εκείνος ο περίπατος, του οποίου δεν επαναλαμβάνεται κανένας κόμβος.

Το μονοπάτι από τον κόμβο i στον κόμβο j ονομάζεται ελάχιστο ή συντομότερο (shortest path) αν οι δύο αυτοί κόμβοι δεν συνδέονται με κανένα άλλο μονοπάτι μικρότερου μήκους. Το μήκος του ελάχιστου μονοπατιού λέγεται αλλιώς και γεωδαισιακή (geodesic) και αντιπροσωπεύει την απόσταση μεταξύ δύο κόμβων. Στο εξής όταν αναφέρεται η απόσταση μεταξύ δυο κόμβων νοείται πάντα η ελάχιστη απόσταση, δηλαδή το συντομότερο μονοπάτι.

2.3 Συνιστώσες- Συνδεσιμότητα

Ένα γράφημα ονομάζεται συνεκτικό ή συνδεδεμένο (connected), αν για οποιουσδήποτε δύο κόμβους του, υπάρχει μια διαδρομή που να τους συνδέει. Στην περίπτωση κατευθυνόμενου γραφήματος, ισχυρά συνεκτικό (strongly connected) μπορεί να θεωρηθεί αν για οποιουσδήποτε κόμβους i,j υπάρχει μια διαδρομή από τον i στον j και μια διαδρομή από τον j στον i κόμβο.

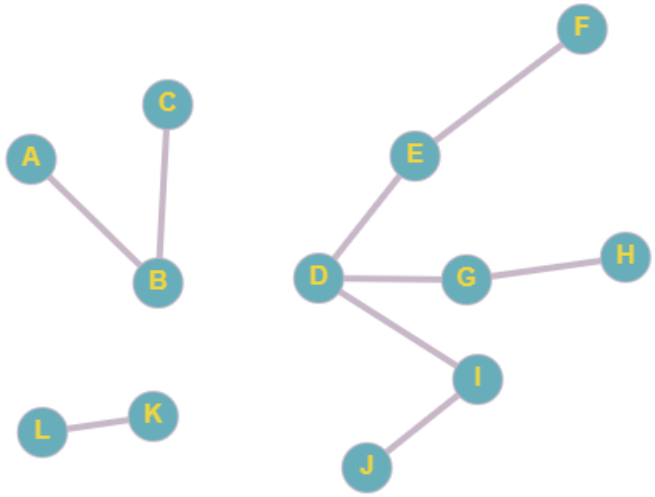
Συνδεδεμένες (ή συνεκτικές) συνιστώσες (connected components) σε ένα μη συνεκτικό γράφο $G = (V, E)$ με $V_1, V_2, V_3, \dots V_k$ μια διαμέριση του συνόλου των κόμβων V σε k υποσύνολα περιγράφονται ως εξής:

- αν ο G είναι μη κατευθυνόμενος γράφος και η διαμέριση είναι τέτοια ώστε δύο κόμβοι i, j που ανήκουν στο ίδιο υποσύνολο κόμβων να συνδέονται με μια

διαδρομή, τότε οι υπογράφοι $G(V_1), G(V_2), \ldots, G(V_k)$ ονομάζονται συνεκτικές συνιστώσες του G.

- αν ο G είναι κατευθυνόμενος γράφος και η διαμέριση είναι τέτοια ώστε δύο κόμβοι i, j που ανήκουν στο ίδιο υποσύνολο κόμβων να συνδέονται με μια διαδρομή από τον κόμβο i στον j, και με μια διαδρομή από τον j στον i, τότε οι υπογράφοι $G(V_1), G(V_2, \ldots, G(V_k)$ ονομάζονται ισχυρά συνδεδεμένες συνιστώσες (strongly connected components) του γράφου G.

Στην εικόνα 2.3 φαίνονται οι τρεις συνδεδεμένες συνιστώσες ενός γραφήματος.



Εικόνα 2.3 : *Συνδεδεμένες συνιστώσες γράφου*

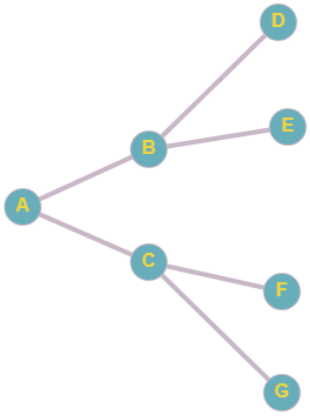
Σε ένα συνδεδεμένο γράφο, η συνεκτικότητα αφορά στον ελάχιστο αριθμό ακμών ή κόμβων που, αν αφαιρεθούν, ο γράφος καθίσταται μη συνδεδεμένος. Διακρίνονται δυο περιπτώσεις:

- η συνεκτικότητα κόμβου (node connectivity), η οποία αντιπροσωπεύει τον ελάχιστο αριθμό κόμβων που πρέπει να αφαιρεθούν για γίνει ο γράφος μη συνεκτικός
- η συνεκτικότητα ακμής (edge connectivity), η οποία αντιπροσωπεύει τον ελάχιστο αριθμό ακμών που πρέπει να αφαιρεθεί για να γίνει ο γράφος μη συνεκτικός. Η ακμή, της οποίας η αφαίρεση οδηγεί σε αποσύνδεση του γράφου ονομάζεται γέφυρα (bridge).

2.4 Δέντρα

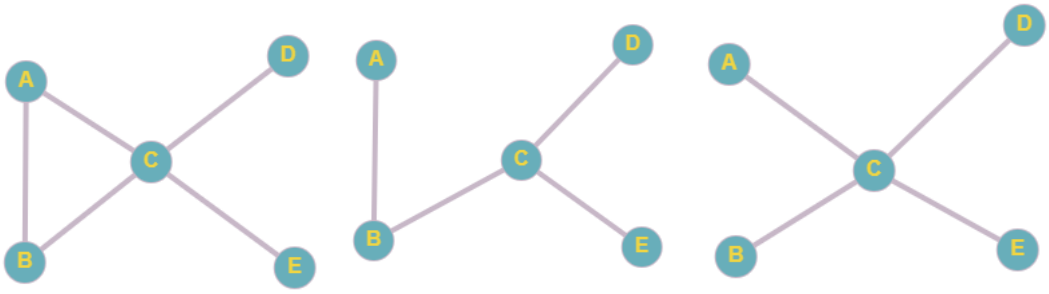
Πριν ορισθεί το δέντρο θα πρέπει να ορισθεί ο κύκλος. Κύκλος (cycle) ή κλειστή διαδρομή σε ένα γράφο $G = (V, E)$ είναι μια ακολουθία κόμβων μήκους l για την οποία ισχύει ότι $v_{i-1}v_i \in E$ για κάθε $i = 1,2, \dots, l-1$ και $v_{i-1}v_0 \in E$.

Δέντρο είναι ο μέγιστος ακυκλικός συνεκτικός γράφος, ή διαφορετικά είναι ο γράφος του οποίου κάθε ακμή είναι γέφυρα. Στα δέντρα υπάρχουν κόμβοι που απέχουν ίδια απόσταση από τη ρίζα. Αυτοί οι κόμβοι ανήκουν στο ίδιο επίπεδο του δέντρου. Στην εικόνα 2.4 φαίνεται ένα δέντρο όπου οι κόμβοι B,C απέχουν απόσταση 1 από τη ρίζα A, ενώ οι κόμβοι D,E,F,G απέχουν απόσταση 2 από τη ρίζα.



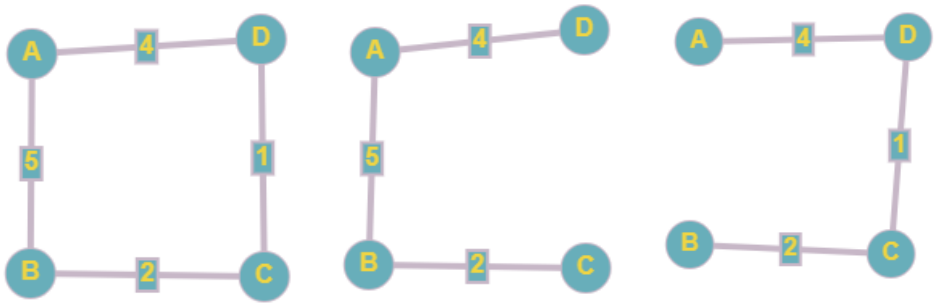
Εικόνα 2.4: *Επίπεδα κόμβων σε δέντρο*

Κάθε συνδεδεμένος γράφος περιέχει ένα (τουλάχιστον) δέντρο επικάλυψης (spanning tree), δηλαδή ένα δέντρο που περιλαμβάνει κάθε κόμβο του γράφου, αλλά ένα υποσύνολο των ακμών του. Στην εικόνα 2.5 φαίνεται ένας γράφος και δυο από τα πιθανά δέντρα επικάλυψης που δημιουργούνται.



Εικόνα 2.5: *Απεικόνιση γράφου 5 κορυφών (αριστερά) και δύο διαφορετικών δέντρων επικάλυψης*

Για κάθε συνδεδεμένο γράφο, υπάρχει επίσης και το δέντρο επικάλυψης ελάχιστου βάρους (minimum weight spanning tree- MST), το οποίο έχει επιπλέον το ελάχιστο άθροισμα βαρών των ακμών του, συγκρινόμενο με οποιοδήποτε άλλο δέντρο επικάλυψης του ίδιου γράφου. Στην εικόνα 2.6 παρουσιάζεται ένας γράφος με βάρη, ένα δέντρο επικάλυψης και το δέντρο επικάλυψης ελάχιστου βάρους.



Εικόνα 2.6: *Weighted Graph (αριστερά), Spanning Tree (κέντρο) και Minimum Spanning Tree (δεξιά)*

Στην ανάλυση σύνθετων δικτύων οι έννοιες που αναφέρθηκαν παραπάνω αποτελούν την απαραίτητη βάση ώστε να γίνουν κατανοητές οι μετρικές και οι μέθοδοι που επιλέχθηκαν στην παρούσα εργασία.

3

Σύνθετα Δίκτυα

Ένα σύνθετο δίκτυο (complex network) μπορεί να ορισθεί ως ένα σύνολο οντοτήτων που αλληλεπιδρούν. Στη κατηγορία αυτή ταξινομούνται τα τεχνολογικά δίκτυα, τα κοινωνικά, τα οικονομικά αλλά και τα βιολογικά δίκτυα. Πρόκειται συνεπώς για κατηγορία δικτύων που καλύπτει ένα ευρύτατο φάσμα εφαρμογών, από τον τομέα του διεθνούς εμπορίου και των διεθνών μεταφορών έως τη διασύνδεση ιστοσελίδων ή τη διάδοση επιδημικών φαινομένων. Αυτού του είδους τα δίκτυα παρουσιάζουν συγκεκριμένες ιδιότητες, όπως είναι η κεντρικότητα των κόμβων ή ο σχηματισμός κοινοτήτων ανάμεσά τους, οι οποίες αναλύονται με τη χρήση διάφορων μετρικών (κεντρικότητα, συντελεστής ομαδοποίησης κ.α.)

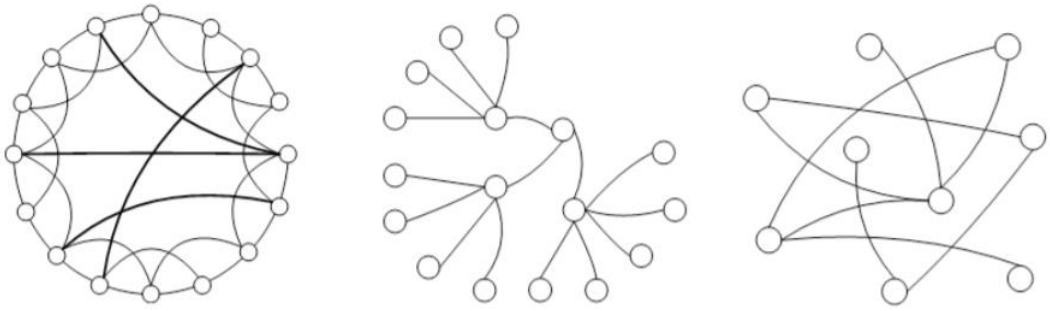
Με άλλα λόγια, τα σύνθετα δίκτυα παρουσιάζουν μεγάλο συντελεστή ομαδοποίησης (clustering coefficient), περιέχουν κοινότητες και περισσότερο ή λιγότερο κεντρικούς κόμβους, χαρακτηριστικά τα οποία απαιτούν την επιστράτευση της θεωρίας γραφημάτων για να περιγραφεί τόσο ο τρόπος δημιουργίας τους, όσο και η ανάπτυξή τους [4].

Για την προσομοίωση και την ανάλυση σύνθετων δικτύων χρησιμοποιούνται κάποια μοντέλα, ανάλογα με τις ιδιότητες των στοιχείων τους. Πιο συγκεκριμένα, διακρίνονται δυο κατηγορίες: α) τα χωρικά μοντέλα και β) τα σχεσιακά μοντέλα. Η πρώτη κατηγορία χρησιμοποιείται για την προσομοίωση δικτύων των οποίων οι κόμβοι συνδέονται βάσει της θέσης τους σε κάποιο γεωμετρικό χώρο, ενώ στη δεύτερη κατηγορία, οι κόμβοι του δικτύου διασυνδέονται βάσει συγκεκριμένων σχέσεων μεταξύ των κόμβων.

Στο κεφάλαιο αυτό παρουσιάζονται συνοπτικά κάποιες βασικές μορφές δικτύων που εμπίπτουν στις δυο προαναφερθείσες κατηγορίες, καθώς και τα βασικά μεγέθη που θα πρέπει να υπολογισθούν για την ανάλυσή τους.

3.1 Τοπολογίες Δικτύων

Στη συγκεκριμένη ενότητα γίνεται αναφορά σε συνθετικά μοντέλα που χρησιμοποιούνται για την προσομοίωση σύνθετων δικτύων, όπως είναι τα Τυχαία Δίκτυα (Random Graphs), τα Δίκτυα Μικρού Κόσμου (Small World Graphs) και τα Δίκτυα Ελεύθερης Κλίμακας (Scale Free Graphs). Στην εικόνα φαίνονται οι τρεις τοπολογίες δικτύων που θα αναλυθούν.



Εικόνα 3.1: *Κατηγορίες δικτύων κατά σειρά: Small World, Scale-free, Random* **[5]**

3.1.1 Τυχαίοι Γράφοι

Στους τυχαίους γράφους (random graphs) οι κόμβοι συνδέονται μεταξύ τους εν γένει με τυχαίο τρόπο. Πρόκειται για σχεσιακά μοντέλα (εφόσον η σύνδεση των κόμβων δεν εξαρτάται από τη θέση τους) στα οποία όλοι οι κόμβοι μπορούν δυνητικά να συνδεθούν μεταξύ τους. Ωστόσο αποτελούν στοιχείο αναφοράς για την αξιολόγηση μοντέλων.

Δυο από τα πιο διαδεδομένα μοντέλα κατασκευής τυχαίων γράφων είναι αυτά των Gilbert και Erdos- Renyi, τα οποία παρουσιάζονται συνοπτικά ακολούθως:

- Το μοντέλο Gilbert $G(n, p)$

Σύμφωνα με την αναφορά [6] ένα τυχαίο δίκτυο δημιουργείται από ένα σύνολο n μεμονωμένων κόμβων, στο οποίο προστίθενται προοδευτικά ακμές με τυχαίο τρόπο. Έτσι κάθε ακμή δημιουργείται με πιθανότητα $p \in (0,1)$, ανεξάρτητα από τις υπόλοιπες. Συνεπώς, ο αναμενόμενος αριθμός ακμών στον γράφο $G(n,p)$ ισούται με $\binom{n}{2}p$. Στην εικόνα 3.2 φαίνεται πώς μεταβάλλεται η μορφή του τυχαίου δικτύου ανάλογα με την πιθανότητα p.

Εικόνα 3.2: *Τυχαίο δίκτυο με διαφορετικές πιθανότητες κατά Gilbert*

- Το μοντέλο Erdos- Renyi $G(n,M)$

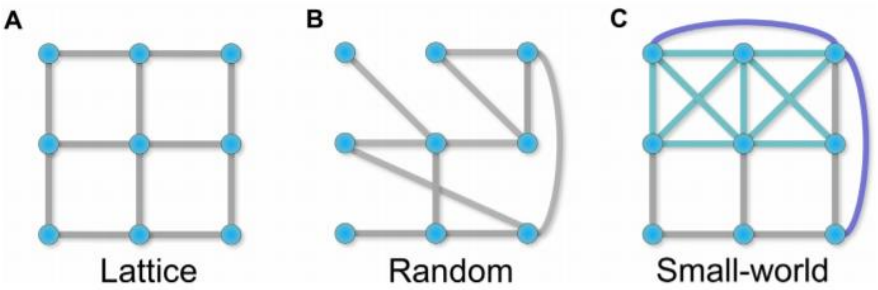
  Οι Erdos και Renyi, [7], θεωρούν πως ο τυχαίος γράφος δημιουργείται με τυχαία επιλογή από το σύνολο των γράφων n κόμβων και M ακμών $G = (n,M)$. Όσο αυξάνεται ο αριθμός των κόμβων και η πιθανότητα δημιουργίας ακμής τείνει στη μονάδα, το μοντέλο του Gilbert $G(n,p)$ γίνεται ισοδύναμο με το $G(n,M)$.

3.1.2 Δίκτυα Μικρού Κόσμου

Ένας εμπειρικός ορισμός για τα δίκτυα Μικρού Κόσμου (Small World Networks) είναι ότι κάθε μεμονωμένος κόμβος μπορεί να συνδεθεί με οποιονδήποτε μη γειτονικό κόμβο του δικτύου, με ένα μικρό αριθμό βημάτων (hops).

Ένα μοντέλο παραγωγής δικτύων Μικρού Κόσμου είναι αυτό των Watts- Strogatz [8], σύμφωνα με το οποίο, σε ένα διατεταγμένο πλέγμα (ordered lattice) n κόμβων, αναδιατάσσονται οι ακμές με πιθανότητα p, έτσι ώστε να συνδέονται μη γειτονικοί κόμβοι, ανεξάρτητα από την αρχική τους απόσταση. Το δίκτυο που παράγεται έχει τη

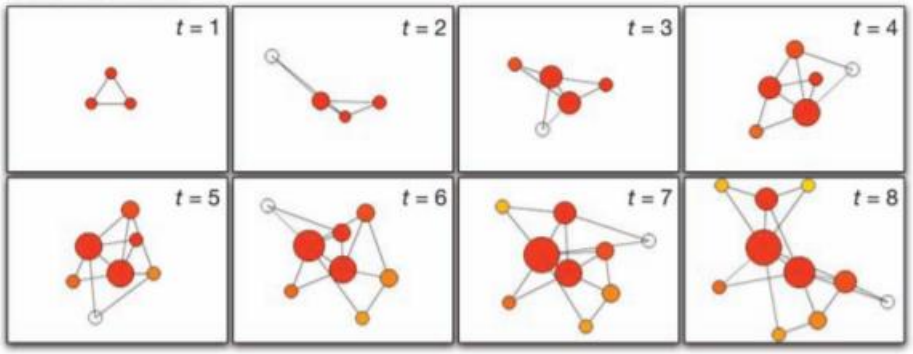
μορφή της εικόνας 3.3. Αξίζει να σημειωθεί ότι όσο η πιθανότητα p αυξάνει, τόσο πιο τυχαία θα δημιουργούνται οι συνδέσεις, άρα το δίκτυο θα τείνει να γίνει τυχαίο.



Εικόνα 3.3: *Η διαφορά των δικτύων Lattice, Random και Small- World* **[9]**

3.1.3 Δίκτυα Ελεύθερης Κλίμακας

Ένα δίκτυο μπορεί να θεωρηθεί ελεύθερης κλίμακας (Scale Free) αν τα χαρακτηριστικά του είναι ανεξάρτητα από το μέγεθός του. Δηλαδή, όσο το δίκτυο αναπτύσσεται, η υποκείμενη γεωμετρία παραμένει ίδια. Ένα δίκτυο ελεύθερης κλίμακας καθορίζεται από την κατανομή των ακμών του, η οποία ακολουθεί το νόμο της δύναμης (power law). Μαθηματικά, η πιθανότητα ένας κόμβος να έχει αριθμό συνδέσεων ίσο με κ, αποτυπώνεται από τη σχέση $P(k) \sim k^{-\gamma}$, όπου $\gamma \in [2,3]$. Αυτό συμβαίνει διότι η κατανομή των ακμών δεν προκύπτει από ανεξάρτητες ποσότητες. Καθώς αυξάνεται το μέγεθος του δικτύου και εισέρχονται νέοι κόμβοι, αυτοί είναι περισσότερο πιθανό να συνδεθούν με εκείνους τους προϋπάρχοντες κόμβους που έχουν ήδη μεγάλο αριθμό συνδέσεων. Αυτό αναφέρεται αλλιώς και προτιμησιακή συνδεσιμότητα (preferential attachment) [10].



Εικόνα 3.4: *Η απλούστερη διαδικασία δημιουργίας τοπολογίας ελεύθερης κλίμακας* **[11]**

Τα δίκτυα ελεύθερης κλίμακας παρατηρούνται σε πολλούς επιστημονικούς τομείς, όπως η τοπολογία των ιστοσελίδων, όπου οι κόμβοι αναπαριστούν τις ιστοσελίδες και οι ακμές αναπαριστούν τις αναφορές (hyperlinks), τα δίκτυα επιστημονικών δημοσιεύσεων, όπου κόμβοι είναι οι δημοσιεύσεις και ακμές είναι οι ετεροαναφορές, τα δίκτυα μεταφορών και πολλά άλλα.

Το μαθηματικό μοντέλο για την κατασκευή δικτύων ελεύθερης κλίμακας προτάθηκε από τους Barabasi και Albert to 1999 [12], σύμφωνα με τους οποίους, ένα τέτοιο δίκτυο προκύπτει αν προστίθενται προοδευτικά κόμβοι σε ένα ήδη υπάρχον δίκτυο, και εισάγονται ακμές έτσι ώστε η πιθανότητα σύνδεσης ενός νέου κόμβου i με έναν προϋπάρχοντα κόμβο j, να είναι ανάλογη του αριθμού των ακμών που ήδη έχει ο j. Αυτό αποτυπώνεται μαθηματικά ως εξής:

$$P(link\ to\ node\ i) = \frac{k_{\mathrm{i}}}{\sum k_j} \qquad \text{(Σχέση 3.1)}$$

Έπειτα από t βήματα το μοντέλο οδηγεί στη δημιουργία τυχαίου δικτύου, ανεπηρέαστου από την κλίμακα, στο οποίο η πιθανότητα ενός κόμβου να έχει k συνδέσεις ακολουθεί το νόμο της δύναμης $P(k) \sim k^{\gamma}$.

3.2 Μετρικές Κοινωνικών Δικτύων

Στην ενότητα αυτή γίνεται αναφορά σε ορισμένες βασικές μετρικές δικτύων, οι οποίες χρησιμοποιούνται στη μελέτη κοινωνικών δικτύων, ώστε να καταστεί δυνατή η αποκάλυψη των ιδιοτήτων και των χαρακτηριστικών τους ή της υποκείμενης δομής και της χρονικής τους εξέλιξης.

3.2.1 Μέσο Μήκος Μονοπατιού

Το μέσο μήκος μονοπατιού (average path length - APL) είναι ο μέσος όρος των βέλτιστων κατά μήκος μονοπατιών, μεταξύ όλων των πιθανών ζευγών κόμβων, και αποτελεί βασική μετρική για την αξιολόγηση της επίδοσης ενός δικτύου.

Σε ένα μη κατευθυνόμενο γράφο $G = (V, E)$, όπου $V = \{v_1, v_2, \dots, v_n)$, η ελάχιστη απόσταση μεταξύ δύο κόμβων v_i , v_j σε βήματα (hops) είναι $d(v_i, v_j)$, και ισούται με μηδέν αν δεν υπάρχει κανένα μονοπάτι (path) που να συνδέει τους δύο κόμβους. Με βάση τα παραπάνω, το μέσο μήκος μονοπατιού δίνεται από τη σχέση:

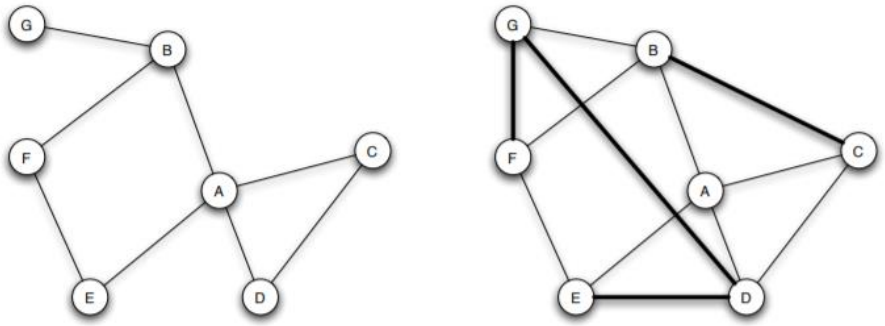
$$APL = \frac{1}{n(n-1)} \sum_{i \neq j} d(v_i, v_j) \qquad \text{(Σχέση 3.2)}$$

Επί της ουσίας, το μέσο μήκος μονοπατιού αντιπροσωπεύει την αναμενόμενη απόσταση μεταξύ δυο κόμβων του δικτύου που επιλέχθηκαν τυχαία. Όσο πιο πυκνό είναι το δίκτυο, τόσο μικρότερο είναι και το APL. Για παράδειγμα, στο γράφο του διαδικτύου το μήκος αυτό είναι πολύ μικρό και αυτό συνεπάγεται πολύ γρήγορη διάδοση της πληροφορίας.

Με βάση τα παραπάνω, σε έναν γράφο που υπάρχουν κοινότητες, δηλαδή περιοχές πυκνής σύνδεσης, αναμένεται μικρό μέσο μήκος μονοπατιού μεταξύ των κόμβων της ίδιας κοινότητας.

3.2.2 Συντελεστής Ομαδοποίησης

Πολύ σημαντική μετρική στη μελέτη κοινωνικών δικτύων είναι ο συντελεστής ομαδοποίησης (clustering coefficient). Είναι μέγεθος που μετρά το βαθμό τριαδικής κλειστότητας (triadic closure) του δικτύου [10], η οποία ορίζει ότι αν δύο κόμβοι έχουν έναν κοινό γείτονα, τότε έχουν μεγάλη πιθανότητα να γίνουν και οι ίδιοι γείτονες, όπως φαίνεται και στην εικόνα 3.5.



***Εικόνα 3.5**: Σχηματισμός νέων συνδέσεων και δημιουργία κλειστών τριγώνων* **[7]**

Στην πράξη ο συντελεστής αυτός αντιπροσωπεύει το βαθμό στον οποίο οι κόμβοι ενός γραφήματος τείνουν να ομαδοποιούνται και να σχηματίζουν κοινότητες. Κυμαίνεται από 0 έως 1 με τις μεγαλύτερες τιμές να δηλώνουν υψηλότερο βαθμό τριαδικής κλειστότητας. Έχουν διατυπωθεί πολλοί ορισμοί για τη συγκεκριμένη μετρική, επισημαίνονται ωστόσο ο τοπικός, ο ολικός και ο μέσος συντελεστής ομαδοποίησης [13] οι οποίοι διατυπώνονται ως ακολούθως:

- Τοπικός Συντελεστής Ομαδοποίησης (Local Clustering Coefficient)

$$C_i = \frac{2\,|\{e_{jk}\}|}{k_i\,(k_i-1)} : \; v_j, v_k \in N_i, e_{jk} \in E \qquad \text{(Σχέση 3.3)}$$

  Ο συγκεκριμένος συντελεστής ποσοτικοποιεί το πόσο κοντά είναι οι γείτονες -του κόμβου για τον οποίο υπολογίζεται- στο να σχηματίσουν πλήρη γράφο (κλίκα). Παίρνει τιμές στο διάστημα [0,1], με μηδενική τιμή αν οι γείτονες του κόμβου δεν συνδέονται μεταξύ τους και μοναδιαία τιμή αν όλοι του οι γείτονες συνδέονται με όλους.

- Ολικός Συντελεστής Ομαδοποίησης (Global Clustering Coefficient)

$$CC_i = \frac{\text{αριθμός ακμών μεταξύ των γειτόνων του } i}{\text{αριθμός όλων των πιθανών ακμών μεταξύ των γειτόνων του } i} \quad \text{(Σχέση 3.4)}$$

- Μέσος Συντελεστής Ομαδοποίησης (Average Clustering Coefficient)

  Υπολογίζεται συνολικά για το δίκτυο από τη σχέση:

$$CC_{net} = \frac{\sum_{i=1}^{n} CC_i}{n} \qquad \text{(Σχέση 3.5)}$$

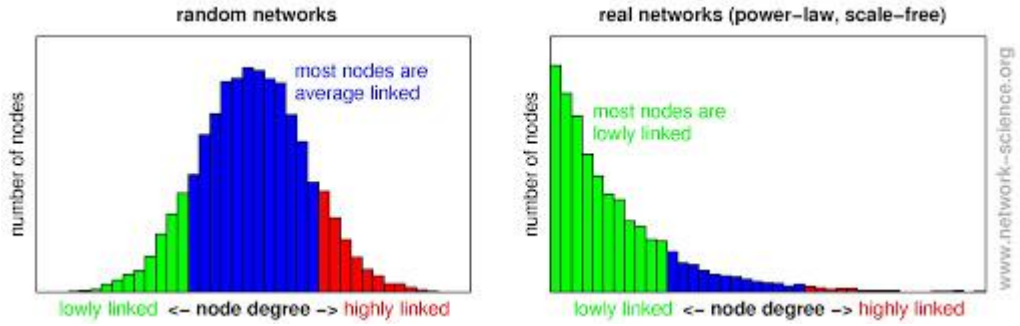
  Πρόκειται για το μέσο όρο των τοπικών συντελεστών ομαδοποίησης όλων των κόμβων του δικτύου.

3.2.3 Κατανομή Βαθμού Κόμβου

Με τον όρο βαθμό κόμβου (node degree) ορίζουμε τον αριθμό των ακμών με τις οποίες ο κόμβος συνδέεται με τους γείτονές του. Στην περίπτωση μη κατευθυνόμενου γραφήματος, διαχωρίζουμε δύο περιπτώσεις: α) τον εσωτερικό βαθμό (in- degree)

που μετρά τις ακμές που φθάνουν στον κόμβο και β) τον εξωτερικό βαθμό (out-degree) που μετρά τις αντίστοιχες που ξεκινούν από τον κόμβο.

Η κατανομή βαθμού κόμβου (degree distribution) δείχνει την πιθανότητα ένας κόμβος να έχει ένα συγκεκριμένο αριθμό γειτόνων. Διαφορετικές κατηγορίες δικτύων παρουσιάζουν διαφορετική κατανομή βαθμού κόμβου [14]. Για παράδειγμα, τα δίκτυα ελεύθερης κλίμακας ως προς το βαθμό ακολουθούν τη κατανομή $P(k) = Ck^{-\gamma}$, ενώ τα τυχαία δίκτυα ακολουθούν την κανονική κατανομή. Στην εικόνα 3.6 φαίνεται η διαφορά στην κατανομή βαθμού κόμβου σε πραγματικά και τυχαία δίκτυα.



Εικόνα 3.6: *Κατανομή βαθμού κόμβου σε τυχαία και πραγματικά δίκτυα* **[15]**

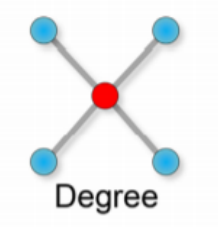
3.2.4 Κεντρικότητα

Η μετρική της κεντρικότητας (Centrality) προσδιορίζει το ρόλο ή τη σημασία ενός κόμβου μέσα σε ένα δίκτυο. Η σημασία ενός κόμβου, δηλαδή το πόσο κεντρικός είναι, μπορεί να έχει διάφορες ερμηνείες, ανάλογα με το δίκτυο ή με το σκοπό για τον οποίο υπολογίζεται. Μπορεί να αφορά μόνο ένα κόμβο (κεντρικότητα βαθμού) ή να συγκρίνει τον κόμβο αυτό με το υπόλοιπο δίκτυο (κεντρικότητα εγγύτητας, ενδιαμεσικότητας). Υπάρχουν συνεπώς πολλές κεντρικότητες οι οποίες απαντούν στην ερώτηση : «Ποιος κόμβος είναι πιο σημαντικός;», και παρουσιάζονται συνοπτικά ακολούθως:

- Κεντρικότητα Βαθμού (Degree Centrality)

  Η κεντρικότητα βαθμού θεωρεί ότι πιο σημαντικός είναι ο εκείνος που συνδέεται με τις περισσότερες ακμές, άρα έχει τους περισσότερους γείτονες. Βασίζεται στην

απλή παραδοχή ότι το πόσο κεντρικός είναι ένας κόμβος είναι συνάρτηση του βαθμού του.



Εικόνα 3.7: *Ο πιο κεντρικός κόμβος ως προς το βαθμό απεικονίζεται με κόκκινο χρώμα* **[9]**

Όσο μεγαλύτερος ο βαθμός ενός κόμβου, τόσο περισσότερο ελέγχει τη ροή πληροφορίας από τη μία άκρη του δικτύου στην άλλη. Έτσι, σε ένα γράφημα $G = (V, E)$ με πίνακα γειτνίασης $A = [a_{ij}]$, η κεντρικότητα βαθμού υπολογίζεται από τη σχέση $C_D(k) = \sum_1^n a_{ik}$. Ωστόσο, η συγκεκριμένη μετρική δεν είναι αντιπροσωπευτική της ροής πληροφορίας, καθώς έχει περισσότερο τοπικό χαρακτήρα (κάθε κόμβος σχετίζεται με τους άμεσους γείτονές του). Για παράδειγμα, μπορεί ένας κόμβος με μεγάλο βαθμό να μην βρίσκεται πάνω στο μονοπάτι από το οποίο διέρχεται πληροφορία, κι έτσι η αφαίρεσή του να μην προκαλέσει σημαντική αλλαγή στη ροή.

- Κεντρικότητα Εγγύτητας (Closeness Centrality)

  Η κεντρικότητα εγγύτητας καθορίζεται από την απόσταση ενός κόμβου από όλους τους υπόλοιπους κόμβους του δικτύου, όπου η απόσταση ορίζεται ως το μήκος του ελάχιστου μονοπατιού (shortest path length) [16]. Αυτό μαθηματικά διατυπώνεται από τη σχέση:

$$C_p(k) = \sum_{i=1}^{n} d(i,k)^{-1} \qquad \text{(Σχέση 3.6)}$$

  Είναι προφανές πως η συγκεκριμένη μετρική αφορά πλήρως συνδεδεμένους γράφους, και καλύπτει το κενό που αφήνει η κεντρικότητα βαθμού ως προς τον έλεγχο της ροής της πληροφορίας. Όσο μεγαλώνει η τιμή της, τόσο πιο «κοντά» είναι ο κόμβος στους υπόλοιπους, άρα και τόσο πιο γρήγορα θα μεταδοθεί η πληροφορία.

- Ενδιαμεσική Κεντρικότητα Κόμβου (Betweenness Centrality)

  Η ενδιαμεσική κεντρικότητα καλύπτει τις περιπτώσεις για τις οποίες δεν μπορεί να υπολογισθεί η κεντρικότητα εγγύτητας, όταν δηλαδή το υπό μελέτη δίκτυο δεν είναι συνεκτικό και έχει μεμονωμένους κόμβους. Η λογική που ακολουθεί είναι ότι ένας κόμβος που βρίσκεται συχνά πάνω στο συντομότερο μονοπάτι που συνδέει δύο άλλους κόμβους του δικτύου, είναι πιο κεντρικός. Σημειώνεται ότι η κεντρικότητα εγγύτητας λαμβάνει υπόψη ότι δεν υπάρχει μόνο μια βέλτιστη διαδρομή μεταξύ δύο κόμβων, και για αυτό περιλαμβάνει τη έννοια της μερικής ενδιαμεσικότητας [17] η οποία ορίζεται ως εξής:

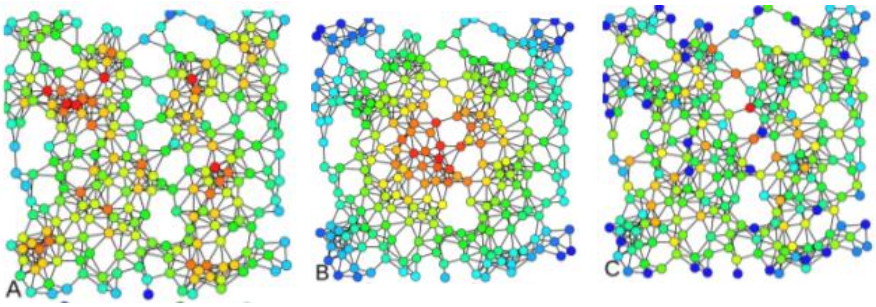
$$b_{ij}(p_k) = \frac{g_{ij}(p_k)}{g_{ij}} \qquad \text{(Σχέση 3.7)}$$

  όπου g_{ij} το σύνολο των συντομότερων μονοπατιών που συνδέουν δύο κόμβους i,j και $g_{ij}(p_k)$ το συντομότερο μονοπάτι που συνδέει τους i,j που όμως περιλαμβάνει τον κόμβο k. Αυτό πρακτικά σημαίνει πως ένας κόμβος που περνά από μερικά μόνο από τα συντομότερα μονοπάτια που συνδέουν δύο άλλους θεωρείται μερικώς κεντρικός.

  Έτσι, η συνολική ενδιαμεσική κεντρικότητα του κόμβου k είναι το άθροισμα όλων των μη διατεταγμένων ζευγών b_{ij} :

$$C_B(p_k) = \sum_{i<j}^{n} \sum_{i<j}^{n} b_{ij}(p_k) \qquad \text{(Σχέση 3.7)}$$

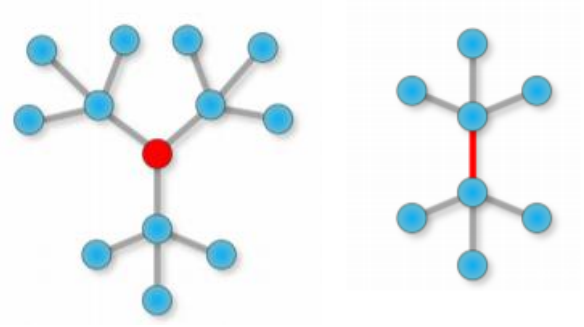
  Στην εικόνα 3.8 φαίνονται η κεντρικότητα βαθμού, εγγύτητας και η ενδιαμεσική. Με κόκκινο χρώμα απεικονίζονται οι κόμβοι με υψηλές τιμές της εκάστοτε εγγύτητας και με μπλε χρώμα οι κόμβοι με χαμηλές τιμές.



Εικόνα 3.8: *Απεικόνιση κεντρικότητας βαθμού (A), εγγύτητας (B) και διαμεσικής (C) στο ίδιο δίκτυο κόμβων* **[45]**

- Κεντρικότητα Ενδιαμεσικότητας Ακμής (Edge Betweenness Centrality)

Η κεντρικότητα ενδιαμεσικότητας ακμής αποτελεί γενίκευση της ενδιαμεσικής κεντρικότητας κόμβου [18] και ισούται με το πλήθος την συντομότερων μονοπατιών που περνούν από μια ακμή του δικτύου. Η ακμή με μεγάλη τιμή της συγκεκριμένης μετρικής εμφανίζει ιδιότητες γέφυρας (bridge-like properties) και αυτό σημαίνει ότι υπάρχει μεγάλη πιθανότητα με την αφαίρεσή της να αλλάξει η ροή στο δίκτυο, ενώ αν η μετρική αυτή σε μια ακμή πάρει τη μέγιστη δυνατή τιμή στο δίκτυο τότε η αφαίρεση της ακμής αυτής θα δημιουργήσει συνιστώσες. Στην εικόνα 3.9 φαίνεται η κεντρικότητα ενδιαμεσικότητας ακμής σε σύγκριση με την ενδιαμεσική κεντρικότητα κόμβου.



Εικόνα 3.9 : *Betweenness Centrality vs Edge Betweenness Centrality* **[9]**

4

Ανάλυση Σύνθετων Δικτύων- Το Παγκόσμιο Δίκτυο Νεοφυών Επιχειρήσεων

Το παγκόσμιο δίκτυο νεοφυών επιχειρήσεων, για τη συγκεκριμένη εφαρμογή, περιλαμβάνει ένα σύνολο από τέτοιου είδους επιχειρήσεις, συνοδευόμενες από κάποια βασικά χαρακτηριστικά-ιδιότητες. Οι οντότητες αυτές συνδέονται μεταξύ τους με άμεσες σχέσεις, όπως είναι οι σχέσεις εξαγοράς, αλλά και με έμμεσες σχέσεις που δημιουργήθηκαν από τα δεδομένα άλλων οντοτήτων που σχετίζονται με αυτές.

Τα δεδομένα που αφορούν στο παγκόσμιο δίκτυο νεοφυών επιχειρήσεων αντλήθηκαν από το www.crunchbase.com και περιλαμβάνουν πρόσωπα, επιχειρήσεις και επενδυτές. Στην πλατφόρμα αυτή τα δεδομένα παρέχονται υπό τη μορφή πινάκων που συνδέονται μεταξύ τους με ένα σχήμα σχεσιακής βάσης δεδομένων.

Οι πίνακες περιλαμβάνουν τις οντότητες (objects) με τα χαρακτηριστικά τους, εξαγορές (acquisitions) που συνδέουν εταιρείες μεταξύ τους, επενδύσεις (investments) και χρηματοδοτήσεις (funding rounds) που συνδέουν χρηματοπιστωτικούς οργανισμούς με εταιρείες και συσχετίσεις (relationships) που συνδέουν άτομα- εργαζόμενους με εταιρείες.

Οι βασικές οντότητες για τις οποίες συλλέγονται τα δεδομένα στην πλατφόρμα και τα χαρακτηριστικά τους (attributes) παρουσιάζονται στον πίνακα 4.1:

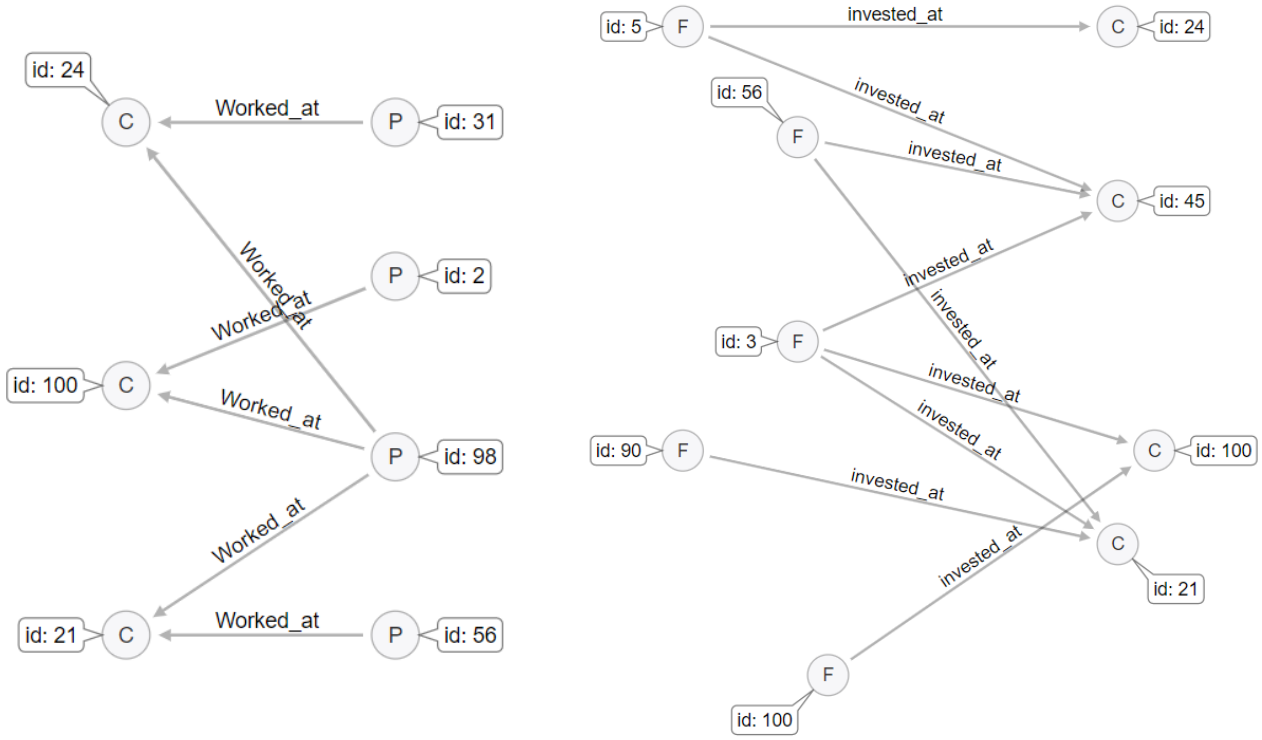
Πίνακας 4.1 Οντότητες, ετικέτες και ιδιότητες.

Entity Name	Label	Attributes
Company	c	id name category_code status founded_at closed_at country_code city funding_rounds funding_total_usd
People	p	id affiliation_name name
Financial Organizations	f	id Name Status country_code state_code city investment_rounds invested_companies

Λόγω του μεγάλου όγκου των δεδομένων έγινε χρήση μη σχεσιακής βάσης δεδομένων, στην οποία αποθηκεύθηκαν 196.553 εταιρείες (c) , 267.694 άτομα (p) και 11.652 χρηματοπιστωτικά ιδρύματα (f) με τα χαρακτηριστικά που φαίνονται στον πίνακα 4.1.

Οι πρώτες συνδέσεις μεταξύ των εταιρειών αντιπροσωπεύουν τις εξαγορές, δηλαδή μια σχέση «μητρικής-θυγατρικής». Οι συνδέσεις αυτές αντλήθηκαν από τα στοιχεία των εξαγορών. Έτσι συνδέθηκαν οι οντότητες με ετικέτα "c".

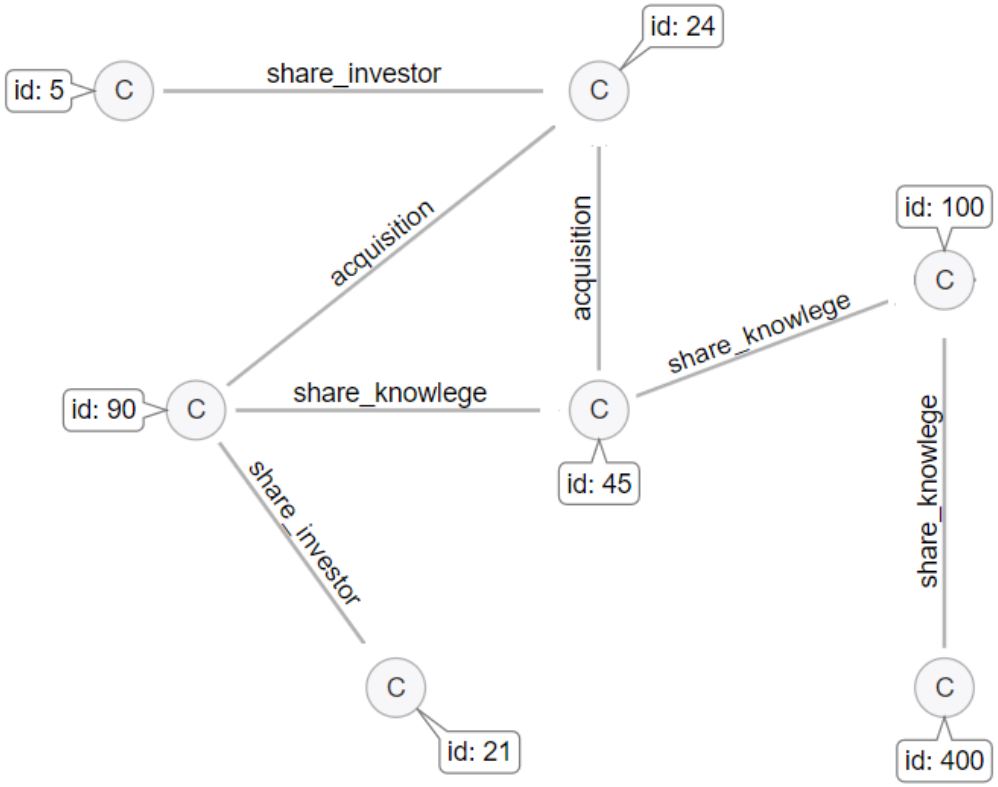
Σε επόμενο στάδιο δημιουργήθηκαν δύο διμερείς γράφοι (bipartite graphs) για τη σύνδεση εταιρειών (c) και ατόμων (p) με βάση τις διασυνδέσεις (relationships), αλλά και για τη σύνδεση χρηματοπιστωτικών οργανισμών (f) και εταιρειών (c) με βάση τις επενδύσεις (investments), πως φαίνεται στην εικόνα 4.1.



Εικόνα 4.1 : *Διμερείς γράφοι εταιρειών-ατόμων (αριστερά), και χρηματοπιστωτικών ιδρυμάτων- εταιρειών (δεξιά).*

Για τη δημιουργία περισσότερων συνδέσεων μεταξύ των κόμβων c του δικτύου έγιναν δύο παραδοχές. Σύμφωνα με την πρώτη, η ροή των εργαζομένων μέσα στο δίκτυο αντιπροσωπεύει τη ροή τεχνογνωσίας. Εταιρείες, δηλαδή στις οποίες έχει εργασθεί το ίδιο άτομο συνδέονται με τη σχέση “share knowledge”.

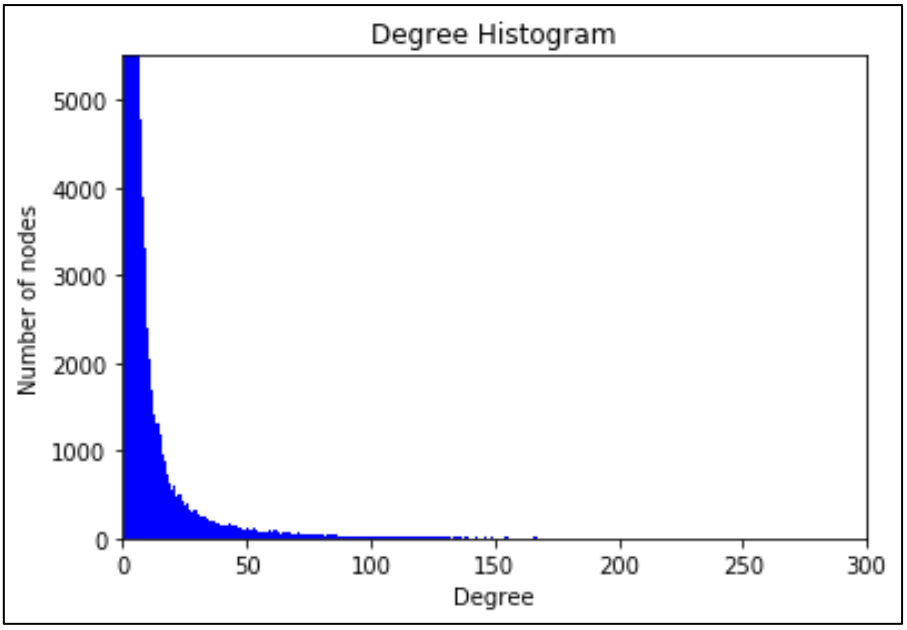
Η δεύτερη παραδοχή συσχετίζει τις εταιρείες με βάση την πηγή χρηματοδότησής τους. Θεωρεί δηλαδή, ότι εταιρείες στις οποίες έχει επενδύσει ο ίδιος χρηματοπιστωτικός οργανισμός βρίσκονται σε παρόμοια κατηγορία επενδυτικού ενδιαφέροντος και συνδέονται με τη σχέση “share investor”. Αυτό πρακτικά σημαίνει ότι, για να προσέλκυσαν τον ίδιο επενδυτή, θα πρέπει να έχουν κάποια κοινά χαρακτηριστικά, προφανή ή και όχι. Στην εικόνα 4.2 φαίνονται όλοι οι τύποι συνδέσεων ανάμεσα στους κόμβους του δικτύου νεοφυών επιχειρήσεων.



Εικόνα 4.2 : *Μέρος παραδείγματος από το Παγκόσμιο Δίκτυο Νεοφυών Επιχειρήσεων*

Στην τελική του μορφή, το παγκόσμιο δίκτυο νεοφυών επιχειρήσεων για τη συγκεκριμένη εφαρμογή περιλαμβάνει 196.553 κόμβους και 539.215 ακμές, και αξιοποιεί πολύ μεγάλη πληροφορία που συνδέεται τόσο άμεσα, όσο και έμμεσα με τις εταιρείες.

Περνώντας από το στάδιο προ-επεξεργασίας (preprocessing), έγινε καθαρισμός τόσο των δεδομένων που συλλέχθηκαν, όσο και του δικτύου, με αφαίρεση των μεμονωμένων κόμβων (orphan nodes) και εξαγωγή της μεγαλύτερης συνδεδεμένης συνιστώσας. Συνεπώς για την ανάλυση, το παγκόσμιο δίκτυο νεοφυών επιχειρήσεων αναπαρίσταται από την μεγαλύτερη συνδεδεμένη συνιστώσα και αποτελείται από 82.300 κόμβους και 500000 ακμές. Ο μέσος βαθμός κόμβου ισούνται με 12,66 και στο γράφημα 4.1 αποτυπώνεται η κατανομή βαθμού κόμβου του δικτύου.



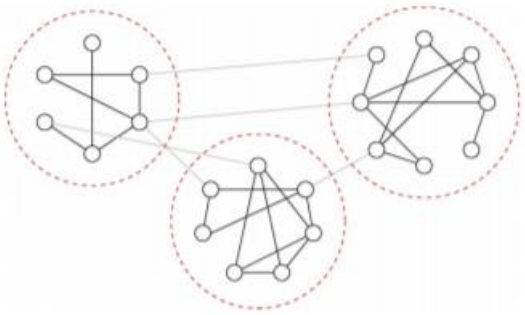
Γράφημα 4.1 : *Κατανομή βαθμού κόμβου για τη μεγαλύτερη συνδεδεμένη συνιστώσα*

Από το γράφημα είναι σαφές πως η κατανομή των ακμών ακολουθεί το νόμο της δύναμης (power-law), χαρακτηριστικό που εμφανίζεται στα δίκτυα ελεύθερης κλίμακας. Παράλληλα, ο μέσος συντελεστής ομαδοποίησης ισούται με $C = 0.6804$, τιμή που συνιστά ένδειξη πως οι κόμβοι του δικτύου τείνουν να σχηματίζουν ομάδες.

5

Ανακάλυψη Κοινοτήτων

Τα πραγματικά δίκτυα δεν κατατάσσονται στην κατηγορία των τυχαίων δικτύων, καθώς εμφανίζουν ιδιαίτερες ιδιότητες και μοτίβα (patterns). Αυτό σημαίνει ότι η υποκείμενη δομή τους και ο τρόπος ανάπτυξής τους δεν μπορεί να προσομοιωθεί με μοντέλα που βασίζονται στην τυχαιότητα. Μια πολύ ενδιαφέρουσα ιδιότητα των δικτύων του πραγματικού κόσμου (real world networks) σε ότι αφορά στη δομή τους, είναι η ανομοιογενής κατανομή των ακμών τους. Αυτή η ιδιότητα οδηγεί σε σχηματισμό κοινοτήτων, δηλαδή ομάδων κόμβων εντός των οποίων η πυκνότητα των ακμών είναι μεγάλη, ενώ ανάμεσα στις ομάδες η πυκνότητα είναι μικρή [19]. Οι κόμβοι που ανήκουν στην ίδια κοινότητα μπορεί να είναι μεταξύ τους περισσότερο όμοιοι ή συνδέονται μεταξύ τους με κάποιον προκαθορισμένο τρόπο [20].



Εικόνα 5.1: *Κοινότητες μέσα σε ένα δίκτυο* **[21]**

Πάνω στην ιδιότητα των σύνθετων δικτύων να σχηματίζουν ομάδες βασίστηκε η διαμέριση γράφων (graph clustering), ένα εργαλείο για την ανακάλυψη κοινοτήτων

μέσα σε ένα δίκτυο με ή χωρίς εκ των προτέρων γνώση του αριθμού τους. Υπάρχουν δύο βασικές προσεγγίσεις για την ανίχνευση κοινοτήτων [20]: i) κατάταξη των κόμβων σε ομάδες συγκρίνοντας τον αριθμό των συνδέσεων εντός της κοινότητας με τον αριθμό των συνδέσεων μεταξύ των κοινοτήτων (density-based clusters) και ii) διαχωρισμό των κόμβων σε ομάδες αναγνωρίζοντας όμοια μοτίβα συνδεσιμότητας (pattern- based clusters). Συνεπώς, διακρίνεται μια σχετικότητα στην ακρίβεια του διαχωρισμού και δεν μπορεί να προσδιορισθεί μια καθολικά αποδεκτή λύση. Στο πλαίσιο αυτό, υπάρχουν πολλές μέθοδοι ανίχνευσης κοινοτήτων σε δίκτυα, οι οποίες κατατάσσονται σε 4 βασικές κατηγορίες [22]:

- **Μέθοδοι με επίκεντρο τον κόμβο (node-centric).**
  Στις μεθόδους αυτές οι κορυφές πρέπει να πληρούν συγκεκριμένα κριτήρια, όπως για παράδειγμα πλήρη αμοιβαιότητα (complete mutuality) και προσβασιμότητα (reachability). Το πρώτο κριτήριο ικανοποιείται μέσα σε μία κλίκα (clique), δηλαδή μια πλήρως συνδεδεμένη ομάδα κόμβων. Αυτή τη λογική ακολουθεί η μέθοδος των Born-Kerbosch [23], στην οποία η ανίχνευση των κοινοτήτων στηρίζεται στην ανίχνευση των μεγιστωτικών (maximal) κλικών. Το δεύτερο κριτήριο πληρείται μέσα σε μία k κλίκα (k-clique), δηλαδή στη μέγιστη συνιστώσα όπου η μεγαλύτερη γεωδαισιακή μεταξύ κόμβων δεν υπερβαίνει τον k αριθμό βημάτων. Με βάση αυτή την προσέγγιση [24] οι κοινότητες προσδιορίζονται μέσω την ανακάλυψης k- κλικών.
- **Μέθοδοι με επίκεντρο την ομάδα (group- centric).**
  Η λογική αυτών των μεθόδων βασίζεται στην απαίτηση τα στοιχεία κάθε κοινότητας να πληρούν κάποιο κριτήριο πυκνότητας. Οι μέθοδοι αυτοί χρησιμοποιούν όλες ή μερικές από τις μεγιστωτικές quasi- κλίκες .
- **Μέθοδοι με επίκεντρο το δίκτυο (network-centric).**
  Σε αυτή την κατηγορία λαμβάνεται υπόψη το σύνολο των συνδέσεων μέσα στο δίκτυο και στόχος είναι να δημιουργηθούν ασύνδετα κομμάτια του αρχικού δικτύου. Παραδείγματα μεθόδων αυτής της λογικής είναι τα μοντέλα ομοιότητας κόμβων (node similarity), τα μοντέλα ελαχιστοποίησης της τομής (cut minimization) και τα μοντέλα μεγιστοποίησης της αρθρωτότητας (modularity maximization). Ένα παράδειγμα μεθόδου με επίκεντρο το δίκτυο είναι και ο αλγόριθμος k-means [25], ο οποίος ανιχνεύει κοινότητες μέσα στο δίκτυο

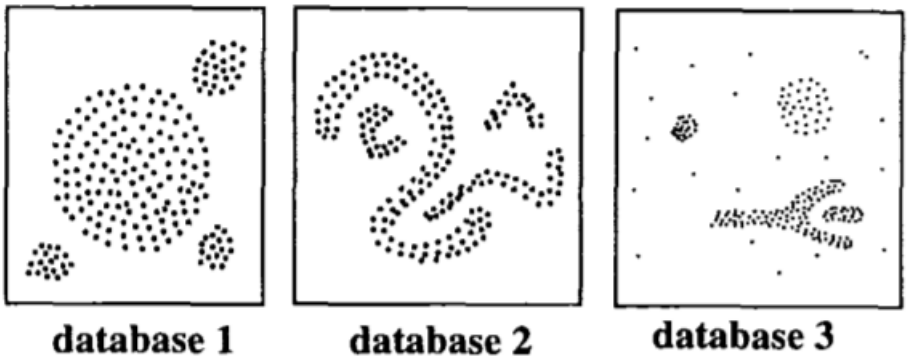
αξιοποιώντας τις συντεταγμένες των κόμβων σε κάποιο γεωμετρικό χώρο, ώστε να υπολογίζει αποστάσεις.

- **Μέθοδοι με επίκεντρο την ιεραρχία (hierarchy- centric).**
  Σε αυτή την ομάδα μεθόδων στόχος είναι η δημιουργία μιας ιεραρχικής δομής κοινοτήτων βασισμένης στην τοπολογία του δικτύου. Εντοπίζονται δύο τύποι ιεραρχικών μεθόδων, ανάλογα με την κατεύθυνση της ανάλυσης (από πάνω προς τα κάτω ή αντίστροφα). Ο πρώτος τύπος ξεκινά από το σύνολο του δικτύου και το διασπά προοδευτικά σε κοινότητες μέχρι να ικανοποιηθεί μια συνθήκη (divisive hierarchical clustering) όπως γίνεται στον αλγόριθμο των Newman και Girvan [21], ενώ ο δεύτερος θεωρεί κάθε μεμονωμένο κόμβο μια ομάδα και συνενώνει διαδοχικά κόμβους στη βάση κάποιου κοινού χαρακτηριστικού, μέχρι να ικανοποιηθεί μια συνθήκη (agglomerative hierarchical clustering). Σε αυτόν το τύπο μεθόδων εντάσσονται τα σχήματα ομαδοποίησης του Stephen Johnson [26].

Στις τέσσερις αυτές κατηγορίες μεθόδων εντάσσεται πληθώρα αλγορίθμων, πέρα από όσες ενδεικτικά αναφέρθηκαν, οι οποίοι απαντούν σε διαφορετικά ερωτήματα και δίνουν διαφορετικά αποτελέσματα στο πρόβλημα της ανίχνευσης κοινοτήτων.

5.1 Ο αλγόριθμος DBSCAN

Ο αλγόριθμος DBSCAN- Density Based Spatial Clustering of Applications with Noise παρουσιάσθηκε το 1996 [27] και είναι μέθοδος διαχωρισμού κοινοτήτων που βασίζεται στην έννοια της πυκνότητας. Ενδείκνυται για διαχείριση μεγάλου όγκου χωρικών δεδομένων και μάλιστα οργανωμένων σε χωρικές βάσεις δεδομένων (Spatial Database Management Systems), καθώς ανακαλύπτει κοινότητες με ακαθόριστη μορφή, και παράλληλα αναγνωρίζει και απομονώνει τον θόρυβο. Στην εικόνα 5.2 απεικονίζονται δείγματα βάσεων δεδομένων, οι κοινότητες που προκύπτουν με κριτήρια πυκνότητας και τα στοιχεία που συνιστούν θόρυβο για τα δεδομένα.



Εικόναν 5.2: *Density- based clustering* **[27]**

Μπορούμε να αναγνωρίσουμε τις κοινότητες στην εικόνα λόγω της υψηλής πυκνότητας που παρουσιάζουν εντός τους. Παράλληλα βλέπουμε μεμονωμένα στοιχεία ανάμεσα στις κοινότητες με σημαντικά χαμηλότερη τιμή πυκνότητας από οποιαδήποτε ομάδα.

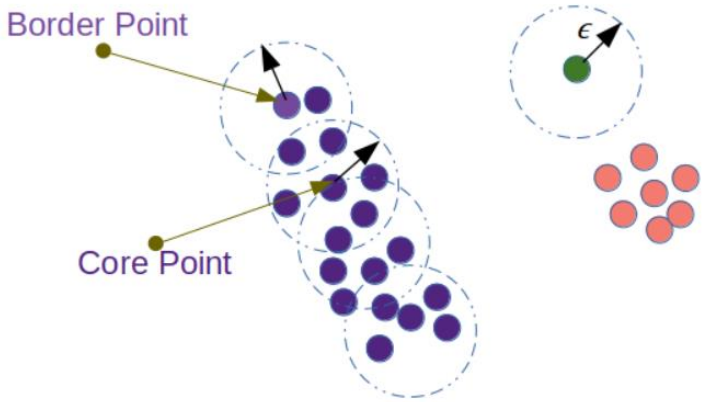
Κεντρική ιδέα του αλγορίθμου είναι ότι για κάθε στοιχείο που ανήκει σε μια κοινότητα υπάρχει ένας ελάχιστος αριθμός στοιχείων γύρω του (στη γειτονιά του), ώστε να πληρείται ένα κριτήριο πυκνότητας. Η γειτονιά κάθε κόμβου καθορίζεται από μια μετρική απόστασης. Σε μια κοινότητα ανιχνεύονται δύο κατηγορίες στοιχείων:

- Πυρήνες (core samples): στοιχεία εντός της κοινότητας που πληρούν το κριτήριο πυκνότητας.
- Μη πυρήνες (non-core samples): στοιχεία που βρίσκονται στο όριο της κοινότητας και δεν πληρούν το κριτήριο πυκνότητας.

Με βάση τα παραπάνω, ο αλγόριθμος προϋποθέτει την εισαγωγή δύο παραμέτρων του ελάχιστου αριθμού σημείων (min_samples) που καθορίζει την πυκνότητα και της ελάχιστης ακτινικής απόστασης (eps).

Κάθε πυρήνας έχει min_sample αριθμό γειτονικών πυρήνων σε απόσταση eps από αυτόν (eps γειτονιά) και επομένως, οι κοινότητες σχηματίζονται αναδρομικά ξεκινώντας από έναν πυρήνα, βρίσκοντας όλους τους γειτονικούς πυρήνες, τους γείτονες των γειτόνων, κ.ο.κ.

Στην εικόνα 5.3 φαίνονται οι πυρήνες και οι μη πυρήνες όταν η παράμετρος min_samples ισούται με 7 και η παράμετρος eps με 1.



Εικόνα 5.3 *Κοινότητες και outliers* **[28]**

Οποιοδήποτε στοιχείο απέχει από πυρήνα απόσταση μεγαλύτερη από eps, αναγνωρίζεται ως ακραία τιμή (outlier) από τον αλγόριθμο.

Η επιλογή της κατάλληλης τιμής για τις δύο παραμέτρους του αλγορίθμου είναι καθοριστικής σημασίας για τον τελικό χωρισμό κοινοτήτων. Πιο συγκεκριμένα, όσο μεγαλύτερη είναι η τιμή min_samples, τόσο μεγαλύτερη είναι η επιθυμητή πυκνότητα για να σχηματισθεί μια κοινότητα. Αντιστοίχως, μικρή τιμή της απόστασης eps θα οδηγήσει σε χαρακτηρισμό των περισσότερων δεδομένων ως θόρυβο, ενώ μια πολύ μεγάλη τιμή απόστασης θα οδηγούσε στην ανίχνευση μια και μόνο κοινότητας που περιλαμβάνει όλα τα δεδομένα.

Ειδικότερα, ο αλγόριθμος διατρέχει όλα τα στοιχεία του dataset ένα προς ένα και εκτελεί τα ακόλουθα βήματα:

i. Αν το στοιχείο έχει ετικέτα πήγαινε στο επόμενο,
ii. βρες τους eps γείτονες του στοιχείου,
iii. αν οι γείτονες είναι λιγότεροι από min_samples τότε βάλε στο στοιχείο την ετικέτα -1 (θόρυβος) και πήγαινε στο επόμενο στοιχείο,
iv. αύξησε το πλήθος της κοινότητας κατά 1 και βάλε τον αριθμό αυτό στο τρέχον στοιχείο, ως ετικέτα,

v. για κάθε ένα από τα eps γειτονικά στοιχεία που δεν έχουν ετικέτα, βάλε την ετικέτα του τρέχοντος και βρες τους γείτονές τους.

Σταματάει όταν δεν υπάρχουν νέα σημεία στη βάση δεδομένων.

Στην εικόνα 5.3 παρουσιάζεται η αλγόριθμος σε μορφή ψευδοκώδικα.

Algorithm 1 The DBSCAN algorithm. Input: A set of points X, distance threshold *eps*, and the minimum number of points required to form a cluster, *minpts*. Output: A set of clusters.

```
1: procedure DBSCAN(X, eps, minpts)
2:     for each unvisited point x ∈ X do
3:         mark x as visited
4:         N ← GETNEIGHBORS(x, eps)
5:         if |N| < minpts then
6:             mark x as noise
7:         else
8:             C ← {x}
9:             for each point x' ∈ N do
10:                N ← N \ x'
11:                if x' is not visited then
12:                    mark x' as visited
13:                    N' ← GETNEIGHBORS(x', eps)
14:                    if |N'| ≥ minpts then
15:                        N ← N ∪ N'
16:                if x' is not yet member of any cluster then
17:                    C ← C ∪ {x'}
```

Εικόνα 5.4: *DBSCAN serial algorithm* **[29]**

Τα πλεονεκτήματα του αλγορίθμου DBSCAN είναι πως ανακαλύπτει κοινότητες ακαθόριστων σχημάτων, προσδιορίζοντας τον θόρυβο και βασίζεται σε κριτήρια πυκνότητας. Στα μειονεκτήματά του κατατάσσεται το γεγονός ότι δεν μπορεί να διαχειρισθεί μεταβολές στην πυκνότητα, είναι δύσκολη η βελτιστοποίηση των παραμέτρων εισόδου και το σημαντικότερο: η σειρά με την οποία εισάγονται τα δεδομένα παίζει ρόλο στην τελική ομαδοποίηση.

5.2 Μεγιστοποίηση της Αρθρωτότητας

Η Αρθρωτότητα (Modularity) είναι μια μετρική που αξιολογεί την διαμέριση ενός δικτύου, υπολογίζει δηλαδή το πόσο «ορθά» έγινε η ομαδοποίηση, λαμβάνοντας υπόψη την κατανομή βαθμού κόμβου.

Η συγκεκριμένη μετρική βασίζεται στην υπόθεση ότι οι κοινότητες που σχηματίζονται μέσα σε ένα δίκτυο θα παρουσιάζουν μεγαλύτερη πυκνότητα εντός τους, παρά μεταξύ τους. Ένας κόμβος, δηλαδή, σχηματίζει περισσότερες συνδέσεις με κόμβους της ίδιας κοινότητας, παρά με αντίστοιχους άλλων ομάδων.

Σε ένα δίκτυο αποτελούμενο από m ακμές, ο αναμενόμενος αριθμός ακμών ανάμεσα σε δύο κόμβους i,j με βαθμό d_i, d_j αντίστοιχα, ισούται με $p = \frac{d_i d_j}{2m}$. Το μέγεθος της αρθρωτότητας για τους κόμβους που ανήκουν σε μια κοινότητα C προκύπτει από τη σχέση 5.1.

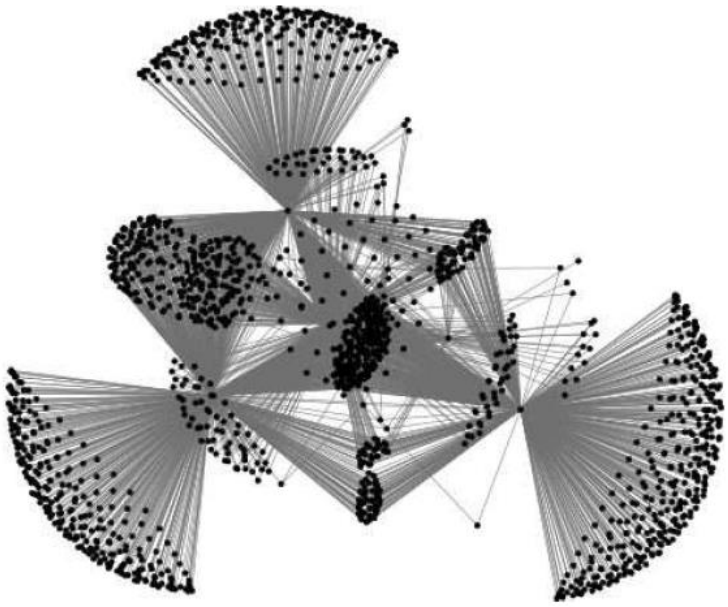
$$Q = \frac{1}{2m} \sum_{l=1}^{k} \sum_{i \in C, j \in C} \left(A_{ij} - \frac{d_i d_j}{2m} \right) \qquad \text{(Σχέση 5.1)}$$

Η μετρική αυτή παίρνει τιμές στο διάστημα [-1,1]. Μια τιμή Q κοντά στο μηδέν δείχνει πως η ομαδοποίηση που έγινε δεν διαφέρει από μια τυχαία διαμέριση. Αντιθέτως, τιμή κοντά στη μονάδα υποδηλώνει καλή διαμέριση.

Η μέθοδος Μεγιστοποίησης της Αρθρωτότητας στοχεύει στον προσδιορισμό του αριθμού κοινοτήτων που μεγιστοποιεί την τιμή του Q. Σύμφωνα την μεθοδολογία που πρότεινε ο Newman [21], όλοι οι κόμβοι αρχικά θεωρούνται ξεχωριστές κοινότητες. Για κάθε ακμή που προστίθεται στο γράφο, ενώνοντας δύο κόμβους σε μια κοινότητα, υπολογίζεται η τιμή του Q. Το βήμα αυτό εκτελείται τόσες φορές, όσοι και οι κόμβοι του δικτύου, εξετάζοντας κάθε φορά αν βελτιώνεται η αρθρωτότητα.

Σε μια προσπάθεια μείωσης του κόστους σε μνήμη και χρόνο, ο Newman μαζί με τους Clauset και Moore πρότειναν την Άπληστη Μεγιστοποίηση της Αρθρωτότητας (Greedy Modularity Maximization) [30], σύμφωνα με την οποία οι κόμβοι αφετηριακά θεωρούνται μεμονωμένες κοινότητες και συνενώνονται επιλέγοντας κάθε φορά τη μεγαλύτερη διαφορά αρθρωτότητας (ΔQ). Μια κοινότητα, δηλαδή, συνδέεται με εκείνη την κοινότητα που θα της δώσει τη μεγαλύτερη αύξηση Q από

όλες τις υπόλοιπες. Στην εικόνα 5.5 φαίνεται η δομή των κοινοτήτων μέγιστης αρθρωτότητας σε ένα δίκτυο.



Εικόνα 5.5: *Οπτικοποίηση της δομής κοινοτήτων μέγιστης αρθρωτοτητας.* **[30]**

Στη συνέχεια στο [31] προτάθηκε μια μέθοδος με ακόμη μικρότερο υπολογιστικό κόστος, όπου συνενώνει τους κόμβους- κοινότητες παράγοντας παράλληλα μια ιεραρχική δομή του δικτύου. Συγκεκριμένα ξεκινά θεωρώντας κάθε κόμβο ως κοινότητα και στη συνέχεια υπολογίζει το ΔQ που θα κέρδιζε μετακινώντας τον στην κοινότητα που ανήκει κάθε γειτονικός του κόμβος. Τελικά επιλέγεται να μετακινηθεί σε εκείνη την κοινότητα για την οποία το ΔQ παίρνει τη μέγιστη θετική τιμή, δημιουργώντας έναν υπερ-κόμβο. Η μέθοδος σταματά όταν σταματήσει να αυξάνεται η τιμή του Q.

Για τη συγκεκριμένη εφαρμογή επιλέχθηκε η προσέγγιση των Clauset-Newman-Moore.

6

Ενσωμάτωση Δικτύων στον Υπερβολικό Χώρο

Οι περισσότερες μέθοδοι ανάλυσης σύνθετων δικτύων υποφέρουν από πολύ μεγάλο υπολογιστικό κόστος. Για την επίλυση του ζητήματος αυτού, η έρευνα στράφηκε σε αποδοτικές λύσεις όπως η κατανεμημένη προσέγγιση στην προσπέλαση δεδομένων και άλλες [32]. Συμπληρωματικά με αυτές τις λύσεις μπορεί να χρησιμοποιηθεί και η ενσωμάτωση γραφημάτων (graph embedding), ώστε να αναπαρασταθεί το δίκτυο σε ένα χώρο χαμηλότερων διατάσεων. Με την ενσωμάτωση του δικτύου (ή μέρους του) σε ένα γεωμετρικό χώρο διαστάσεων d μικρότερων του αρχικού, αποδίδονται σε κάθε κόμβο συντεταγμένες του χώρου αυτού, με τρόπο ώστε να διατηρείται όσο γίνεται η δομή του δικτύου και οι ιδιότητες των κόμβων. Αυτό έχει ως αποτέλεσμα το δίκτυο να αναπαρασταθεί από ένα σύνολο διανυσμάτων χαμηλών διαστάσεων και έτσι οι αλγόριθμοι γραφημάτων (όπως π.χ. η εύρεση του ελάχιστου μονοπατιού) να είναι περισσότερο αποδοτικοί.

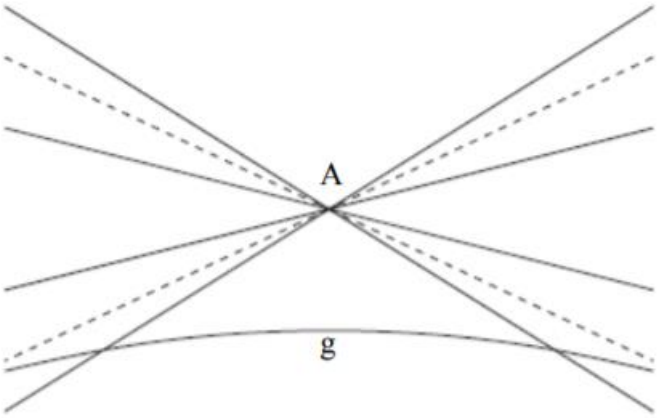
Έχουν προταθεί αρκετά συστήματα ενσωμάτωσης γραφημάτων τα οποία να δίνουν αξιόπιστη αναπαράσταση του δικτύου, χωρίς μεγάλη απώλεια πληροφορίας. Κάποια εφαρμόζουν ενσωμάτωση στον ευκλείδειο χώρο όπως το Orion [33], άλλα αξιοποιούν τον σφαιρικό χώρο [34] για εφαρμογές όρασης υπολογιστών, και υπάρχουν και περιπτώσεις που χρησιμοποιείται ο υπερβολικός χώρος [35], με μια σειρά αλγορίθμων όπως αναφέρονται στα [36] και [37].

6.1 Έννοιες της Υπερβολικής Γεωμετρίας

Η υπερβολική γεωμετρία ή γεωμετρία του Lobachevsky είναι μια μη ευκλείδεια γεωμετρία, στη οποία δεν ισχύει το $5^{ο}$ αξίωμα του Ευκλείδη που αφορά στις παράλληλες ευθείες. Πιο συγκεκριμένα, αντί του αξιώματος των παράλληλων ισχύει το ακόλουθο:

«Σε επίπεδο δύο διαστάσεων, για κάθε ευθεία *ε* και σημείο Ο εκτός της *ε*, υπάρχουν άπειρες ευθείες που διέρχονται από το Ο και δεν τέμνουν την *ε*»

Ένα άλλο γνώρισμα της υπερβολικής γεωμετρίας είναι ότι η απόσταση μεταξύ υπερπαράλληλων ευθειών (δηλαδή ευθειών που δεν τέμνονται) μεγαλώνει όταν αυτές εκτείνονται στο άπειρο.



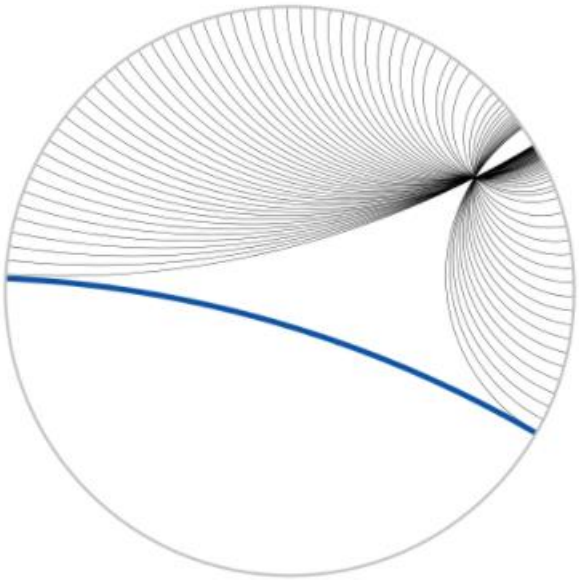
Εικόνα 6.1: *Απεικόνιση παράλληλων της g που διέρχονται από το A*

Υπάρχουν πέντε (5) μοντέλα υπερβολικών χώρων, κάθε ένα από τα οποία επικεντρώνεται σε διαφορετικά στοιχεία της υπερβολικής γεωμετρίας [38], ανάλογα με την εφαρμογή. Η σημαντικότερη ιδιότητα των υπερβολικών χώρων, που τους καθιστούν κατάλληλους για την ενσωμάτωση σύνθετων δικτύων, είναι ότι επεκτείνονται εκθετικά, άρα γρηγορότερα, σε σύγκριση με τους ευκλείδειους χώρους, ενώ παράλληλα περιλαμβάνουν περισσότερη πληροφορία σε μικρότερο χώρο.

Στη συνέχεια αναλύονται δυο από τα πέντε μοντέλα μοντέλα υπερβολικών χώρων, ο δίσκος Poincaré και το Υπερβολοειδές.

6.1.1 Το μοντέλο του δίσκου του Poincaré (P.D.M.)

Το μοντέλο του δίσκου του Poincaré συνιστά τροποποίηση του μοντέλου του Klein [39] και παρέχει διαισθητικά καλύτερη αντίληψη του υπερβολικού χώρου. Έχει ως επίπεδο ένα μοναδιαίο δίσκο και ως ευθείες είτε τα τόξα κύκλων κάθετων στη περιφέρεια του δίσκου, είτε τις διαμέτρους του δίσκου.



Εικόνα 6.2: *Παράλληλες γραμμές στο P.D.M.* **[40]**

Η απόσταση στο P.D.M. είναι πολύ μεγαλύτερη από την ευκλείδεια κοντά στην περιφέρεια του δίσκου, ενώ είναι πρακτικά ίση με την ευκλείδεια για σημεία πολύ κοντά στο κέντρο. Οποιοδήποτε τόξο ευκλείδειου κύκλου- δηλαδή ευθεία στο δίσκο του Poincaré, όσο πλησιάζει στην περιφέρεια θα έχει πολλαπλάσια μεγαλύτερο μήκος από το ευκλείδειο. Με βάση την παραπάνω συλλογιστική, η περιφέρεια του δίσκου ονομάζεται σύνορο στο άπειρο (infinite edge), δεν ανήκει στον υπερβολικό χώρο και οι υπερβολικές ευθείες επεκτείνονται στο άπειρο. Επίσης, οι ευθείες στο δίσκο είναι κάθετες στην περιφέρεια.

Ο μαθηματικός τύπος της απόστασης δύο σημείων x_1, x_2 στο P.D.M. ισούται με:

$$d_{x_1x_2} = \text{atanh}(\frac{|x_1-x_2|}{1-x_1\overline{x_2}}) \qquad \text{(Σχέση 6.1)}$$

6.1.2 Το μοντέλο του Υπερβολοειδούς (Hyperboloid Model)

Το μοντέλο του Υπερβολοειδούς βασίζεται σε ένα υπερβολοειδές n διαστάσεων στον (n+1) χώρο Minkowski $\mathbb{R}^{n:1}$. Το υπερβολοειδές n διαστάσεων είναι ο γεωμετρικός τόπος σημείων όπου ισχύουν τα εξής [41]:

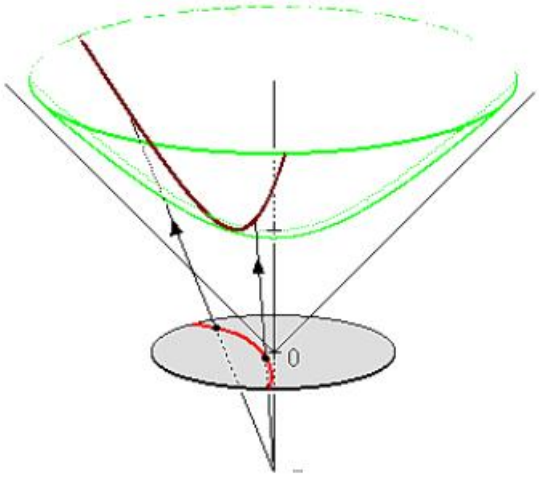
$$\mathbb{H}^n = \{\, x \in \mathbb{R}^{n:1} \,|\langle x, x\rangle_{n:1} = -1, x_{n+1} > 0 \,\} \qquad \text{(Σχέση 6.2)}$$

Όπου $\langle u, v\rangle_{n:1} = \sum_{i=1}^{n} u_i v_i - u_{n+1}\, v_{n+1}$

Η απόσταση μεταξύ δύο σημείων u,v στο υπερβολοειδές δίνεται από τη σχέση:

$$d_{\mathbb{H}^n}(u, v) = \operatorname{arccosh}(-\langle u, v\rangle_{n:1}), u, v \in \mathbb{H}^n \qquad \text{(Σχέση 6.3)}$$

Στην εικόνα 6.3 φαίνεται ο δίσκος του Poincaré και το υπερβολοειδές με πράσινο χρώμα.



Εικόνα 6.3: *Δίσκος του Poincaré και Υπερβολοειδές* **[42]**

Για τον υπολογισμό της υπερβολικής απόστασης μεταξύ δύο σημείων μπορεί να χρησιμοποιηθεί ο ακόλουθος τύπος [35]:

$$d(u, v) = arccosh(\sqrt{(1 + \sum_{i=1}^{n} u_i^2)(1 + \sum_{i=1}^{n} v_i^2)} - \sum_{i=1}^{n} u_i v_i) \times |c| \qquad \text{(Σχέση 6.4)}$$

όπου c η κυρτότητα του χώρου.

6.2 Η καταλληλότητα του υπερβολικού χώρου για την ενσωμάτωση σύνθετων δικτύων

Συγκρινόμενοι με τους ευκλείδειους χώρους, οι υπερβολικοί χώροι κλιμακώνονται εκθετικά ως προς την ακτινική συντεταγμένη. Πιο συγκεκριμένα, στο δίσκο του Poincaré που αναλύθηκε παραπάνω, και για καμπυλότητα ίση με -1, η περιφέρεια C και το εμβαδόν A κύκλου ακτίνας r δίνεται από τις σχέσεις:

$$C(r) = 2\pi \sinh(r) \qquad \text{(Σχέση 6.5)}$$

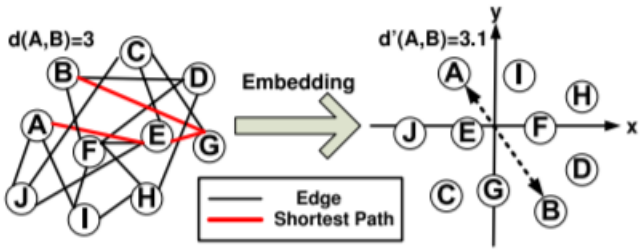
$$A(r) = 4\pi sinh^2(\frac{r}{2}) \qquad \text{(Σχέση 6.6)}$$

Για μικρή ακτίνα r, επομένως, ισχύει $C(r) \cong 2\pi r$ και $A(r) \cong \pi r^2$. Όσο μεγαλώνει η ακτίνα, όμως, και δεδομένου ότι $\sin h(r) = \frac{e^r - e^{-r}}{2}$, η περιφέρεια και το εμβαδόν μεγαλώνουν ασυμπτωτικά σε σχέση με την ακτίνα [38].

Η ιδιότητα της εκθετικής κλιμάκωσης, καθιστά τον υπερβολικό χώρο εξαιρετικά κατάλληλο για την ενσωμάτωση μεγάλου όγκου δεδομένων, όπως συνήθως συμβαίνει με τα σύνθετα δίκτυα. Επιπλέον, υπάρχει και η θεωρία ότι κάθε σύνθετο δίκτυο υποκρύπτει την υπερβολική γεωμετρία [38].

6.3 Η ενσωμάτωση Rigel

Κατά την ενσωμάτωση Rigel [33], κάθε κόμβος λαμβάνει συντεταγμένες με τρόπο ώστε η απόστασή του από οποιονδήποτε άλλο κόμβο στο μετρικό χώρο να είναι σχεδόν ίση με το αντίστοιχο μήκος του συντομότερου μονοπατιού μεταξύ των κόμβων στο αρχικό δίκτυο. Στην εικόνα 6.4 φαίνεται ότι η απόσταση των κόμβων A,B στο αρχικό δίκτυο είναι ίση με 3 (συντομότερο μονοπάτι), ενώ στον ενσωματωμένο γράφο η απόστασή τους υπολογίζεται 3,1.



Εικόνα 6.4: *Ενσωμάτωση στον ευκλείδειο χώρο. Απόσταση κόμβων A,B : d(A,b)=3, ευκλείδεια απόσταση κόμβων A,B : 3,1.* **[35]**

Για την ενσωμάτωση ενός δικτύου n κόμβων με τη μέθοδο Rigel, αρχικά επιλέγεται μικρός αριθμός κόμβων που θα χρησιμοποιηθούν ως σημεία αναφοράς (landmarks). Υπολογίζονται οι αποστάσεις όλων των κόμβων του δικτύου από τους κόμβους αναφοράς και στη συνέχεια οι τελευταίοι ενσωματώνονται στον μετρικό χώρο με τρόπο ώστε να διατηρείται κατά το δυνατόν η μεταξύ τους απόσταση κοντά στο μήκος του συντομότερου μονοπατιού που τους συνδέει. Σε δεύτερη φάση ενσωματώνονται όλοι οι υπόλοιποι κόμβοι του δικτύου με βάση κάποιο υποσύνολο από τα landmarks και την απόστασή τους από αυτά. Οι συντεταγμένες των λοιπών κόμβων υπολογίζονται ως λύση ενός προβλήματος γραμμικού προγραμματισμού που βελτιστοποιεί τις αποστάσεις τους από τα σημεία αναφοράς.

7

Ανακάλυψη κοινοτήτων στο παγκόσμιο δίκτυο νεοφυών επιχειρήσεων

Λαμβάνοντας υπόψη τη σχετικά μεγάλη τιμή του μέσου συντελεστή ομαδοποίησης στο υπό μελέτη δίκτυο ($C = 0.6804$), στο παρόν κεφάλαιο εφαρμόζονται δύο μέθοδοι για την ανακάλυψη κοινοτήτων εντός του. Ειδικότερα παρουσιάζεται μια παραλλαγή της μεθόδου DBSCAN ως προς τη μετρική της απόστασης που χρησιμοποιεί, και επίσης επιχειρείται ομαδοποίηση με μια μέθοδο μεγιστοποίησης της αρθρωτότητας.

7.1 Ενσωμάτωση δικτύου στον υπερβολικό χώρο

Όπως έχει ήδη αναφερθεί, μελέτες έχουν δείξει πως η υπερβολική γεωμετρία ενυπάρχει σε κάθε σύνθετο δίκτυο, ενώ παράλληλα οι υπερβολικοί μετρικοί χώροι ενσωματώνουν μεγάλο όγκο πληροφορίας σε μικρότερο χώρο [43]. Ακολουθώντας αυτή τη λογική, το δίκτυο των επιχειρήσεων ενσωματώθηκε στον υπερβολικό χώρο μέσω της ενσωμάτωσης Rigel [44]. Αρχικά επιλέχθηκε ένας αριθμός κόμβων-ορόσημων (Landmarks) τα οποία ενσωματώθηκαν πρώτα. Όλοι οι υπόλοιποι κόμβοι έλαβαν συντεταγμένες βάσει της απόστασής τους από τα ορόσημα. Τα ορόσημα δεν επιλέχθηκαν με τυχαίο τρόπο από το δίκτυο, αλλά με τρόπο ώστε να απλοποιούνται κατά το δυνατόν οι μετέπειτα υπολογισμοί. Έτσι, υποψήφιοι κόμβοι-ορόσημα ήταν εκείνοι με τη μεγαλύτερη κεντρικότητα βαθμού κόμβου.

Για την ενσωμάτωση έπρεπε να προσδιορισθεί η τιμή δύο παραμέτρων: αριθμός ορόσημων (L) και αριθμός συντεταγμένων (b), οι οποίες και προσδιορίσθηκαν πειραματικά. Δεδομένου ότι καλύτερη ενσωμάτωση είναι εκείνη που δίνει υπερβολική απόσταση μεταξύ δύο κόμβων περίπου ίση με την πραγματική απόσταση (σε βήματα) των ίδιων κόμβων στο δίκτυο, ελέγχθηκαν δειγματοληπτικά 1000 ζεύγη κόμβων του δικτύου και η τελική επιλογή των παραμέτρων καθορίσθηκε από το μέσο τετραγωνικό σφάλμα των αποστάσεων για τα ζεύγη αυτά. Σημειώνεται ότι επιλέχθηκε μικρός αριθμός ορόσημων ως προς το μέγεθος του δικτύου, για να ολοκληρωθεί η ενσωμάτωση σε εύλογο χρόνο. Στον πίνακα 7.1 παρουσιάζονται συνοπτικά τα αποτελέσματα των δοκιμών.

Πίνακας 7.1: Αποτελέσματα πειραμάτων για τις παραμέτρους ενσωμάτωσης

curvature	**# landmarks**	**# primary landmarks**	**# dimensions**	**RMSE**
-1	50	50	2	2,5441
-1	100	20	2	2,0236
-1	100	20	3	1,7763
-1	100	30	2	2,1079
-1	100	30	3	1,7530
-1	100	30	5	1,5327
-1	100	50	2	1,2380
-1	100	50	3	1,7486
-1	100	50	5	1,5421
-1	100	50	6	1,4930
-1	100	100	2	2,6626

Τα πειράματα οδήγησαν σε έναν υπερβολικό χώρο έξι διαστάσεων με καμπυλότητα ίση με -1 και χρήση 100 ορόσημων για την ενσωμάτωση 82.300 κόμβων.

7.2 Εφαρμογή του αλγορίθμου DBSCAN με υπερβολικές αποστάσεις.

Ο αλγόριθμος DBSCAN βασίζει την ομαδοποίηση σημείων στο χώρο στη μέτρηση της «πυκνότητας» κόμβων γύρω από κάθε σημείο προς ομαδοποίηση. Στο πλαίσιο της εργασίας, αντί των προκαθορισμένων μετρικών της μεθόδου, δημιουργήθηκε μια νέα συνάρτηση που υπολογίζει την υπερβολική απόσταση μεταξύ δύο οποιωνδήποτε κόμβων, βασισμένη στις συντεταγμένες που τους έχουν αποδοθεί στον υπερβολικό χώρο. Έτσι, ο πίνακας των υπερβολικών συντεταγμένων και η συνάρτηση hyperbolic_distance που κατασκευάσθηκε, τροφοδοτήθηκαν στο μοντέλο.

Στη συνέχεια έπρεπε να καθορισθούν οι παράμετροι epsilon και minimum_samples, δηλαδή οι παράμετροι ακτίνας απόστασης και τα κριτήρια πυκνότητας για το διαχωρισμό. Αυτές οι παράμετροι είναι πολύ σημαντικές για την ποιότητα του διαχωρισμού σε κοινότητες, καθώς πολύ μικρή τιμή ακτίνας θα χαρακτήριζε τους περισσότερους κόμβους ως θόρυβο, ενώ πολύ μεγάλη αντίστοιχη τιμή θα κατέτασσε όλους τους κόμβους στην ίδια κοινότητα. Αντιστοίχως, μεγάλη τιμή min_samples θα μεγάλωνε τόσο το κριτήριο πυκνότητας για τη σύσταση ομάδας, με αποτέλεσμα να ανακαλύψει μικρό αριθμό ομάδων, κ.ο.κ.

Καθώς το αρχικό δίκτυο των επιχειρήσεων αποτελείται από πολλούς κόμβους, είναι αδύνατη η διατήρηση στην κύρια μνήμη ενός πίνακα που θα διατηρούσε την πληροφορία για τις μεταξύ τους αποστάσεις. Συνεπώς, αυξάνεται η υπολογιστική πολυπλοκότητα του αλγορίθμου DBSCAN εφόσον σε κάθε επανάληψη χρειάζεται ο επανυπολογισμός όλων των αποστάσεων. Σαν μια λύση για τον πειραματικό προσδιορισμό των παραμέτρων eps και min_samples, χρησιμοποιήθηκε ένα κατάλληλο υποδίκτυο της τάξης των 5000 κόμβων. Οι δοκιμές, όπως φαίνεται και στον πίνακα 7.2, συνέστησαν την εφαρμογή του μοντέλου με τις δύο παραμέτρους να ισούνται με 2.

Πίνακας 7.2: Αριθμός παραγόμενων ομάδων και θορύβου για διαφορετικές τιμές eps και min_samples

epsilon	min_samples	# clusters	# noise points
0.4	2	4	5397
0.5	2	16	5372
0.6	3	3	5395

0.6	2	29	5343
0.7	2	53	5287
0.8	2	106	5147
1	2	217	4736
2	**2**	**28**	**1392**
3	2	2	192
4	2	2	20
5	2	1	4
4	3	2	20
4	4	2	20
4	5	1	24

7.3 Εφαρμογή του αλγορίθμου Greedy Modularity Communities

Για την παρούσα εφαρμογή επιλέχθηκε η μέθοδος Άπληστης Μεγιστοποίησης της Αρθρωτότητας των Clauset και Moore, όπου η συνένωση των κόμβων σε κοινότητες γίνεται με κριτήριο τη μεγιστοποίηση της διαφοράς αρθρωτότητας (ΔQ). Η επιλογή βασίστηκε στο γεγονός ότι η συγκεκριμένη μέθοδος μπορεί να διαχειρισθεί μεγάλα δίκτυα αποδοτικά ως προς το κόστος. Συνεπώς, ο αλγόριθμος Greedy Modularity Communities εφαρμόσθηκε στη μεγαλύτερη συνδεδεμένη συνιστώσα και εξήγαγε 1697 κοινότητες. Η τιμή της αρθρωτότητας για τη συγκεκριμένη διαμέριση ισούται με $Q = 0.51346$, γεγονός που υποδεικνύει, σε συνδυασμό με την υψηλή τιμή του συντελεστή ομαδοποίησης ότι το παραχθέν δίκτυο επιχειρήσεων παρουσιάζει δομή οργανωμένη σε κοινότητες.

8

Αποτελέσματα

Στο παρόν κεφάλαιο παρουσιάζονται τα αποτελέσματα της εφαρμογής των δυο μεθόδων ομαδοποίησης στο δίκτυο νεοφυών επιχειρήσεων. Αρχικά παρατίθενται τα αποτελέσματα της ομαδοποίησης με τον αλγόριθμο DBSCAN στο ενσωματωμένο δίκτυο επιχειρήσεων και στη συνέχεια παρουσιάζονται τα αποτελέσματα της ομαδοποίησης με κριτήριο τη μεγιστοποίηση της αρθρωτότητας. Οι κοινότητες που προέκυψαν αξιολογούνται με στατιστικά εργαλεία και μετρικές ανάλυσης δικτύων. Οι αλγόριθμοι υλοποιήθηκαν στη γλώσσα προγραμματισμού Python, ενώ η ανάκτηση και διαχείριση των δεδομένων πραγματοποιήθηκε με τη χρήση της μη σχεσιακής βάσης δεδομένων Neo4j και γλώσσα Cypher Query Language.

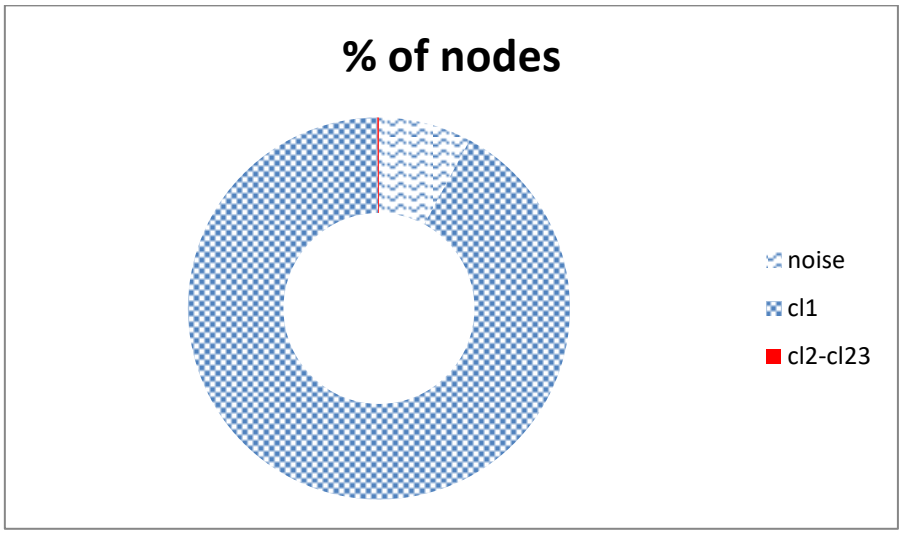
8.1 Αποτελέσματα DBSCAN

Ο αλγόριθμος DBSCAN εφαρμόσθηκε με παραμέτρους epsilon=2, minimum samples=2 και metric=hyperbolic distance. Ολοκληρώθηκε σε 28 ώρες και στον πίνακα 8.1 φαίνονται συγκεντρωτικά τα αποτελέσματα.

Πίνακας 8.1: Αποτελέσματα DBSCAN

number of nodes	number of edged	estimated number of clusters	estimated number of noise points	number of nodes in the largest cluster
82.360	521.172	23	6.484	75.774

Παρατηρήθηκε συσσώρευση άνω του 90% των κόμβων στην ίδια κοινότητα, ενώ ποσοστό λίγο μικρότερο του 8% έμεινε αταξινόμητο και χαρακτηρίστηκε από τον αλγόριθμο ως θόρυβος, όπως διαπιστώνεται και στο γράφημα 8.1.



Γράφημα 8.1 *Ποσοστό κόμβων στις 3 βασικές κατηγορίες: 92% cl 1, 7.9% noise, 0.1 % cl 2-23*

Εφόσον σχεδόν το σύνολο των κόμβων ταξινομήθηκαν σε μια κατηγορία, η διαμέριση που πραγματοποίησε ο συγκεκριμένος αλγόριθμος δεν ενδείκνυται για την εξαγωγή ασφαλούς συμπεράσματος στην παρούσα εφαρμογή.

8.2 Αποτελέσματα Greedy Modularity Communities

Ο αλγόριθμος Greedy Modularity Communities εφαρμόσθηκε στην ίδια συνιστώσα και εντόπισε 1679 κοινότητες εκ των οποίων οι 10 μεγαλύτερες συγκέντρωσαν περίπου το 75% των κόμβων, όπως φαίνεται και στον πίνακα 8.2.

Πίνακας 8.2: Αριθμός και ποσοστό κόμβων ανά ομάδα σύμφωνα με τον GMC

	number of nodes	**%**
Cluster 1	17.167	20,84
Cluster 2	17.250	20,95
Cluster 3	11.521	13,99

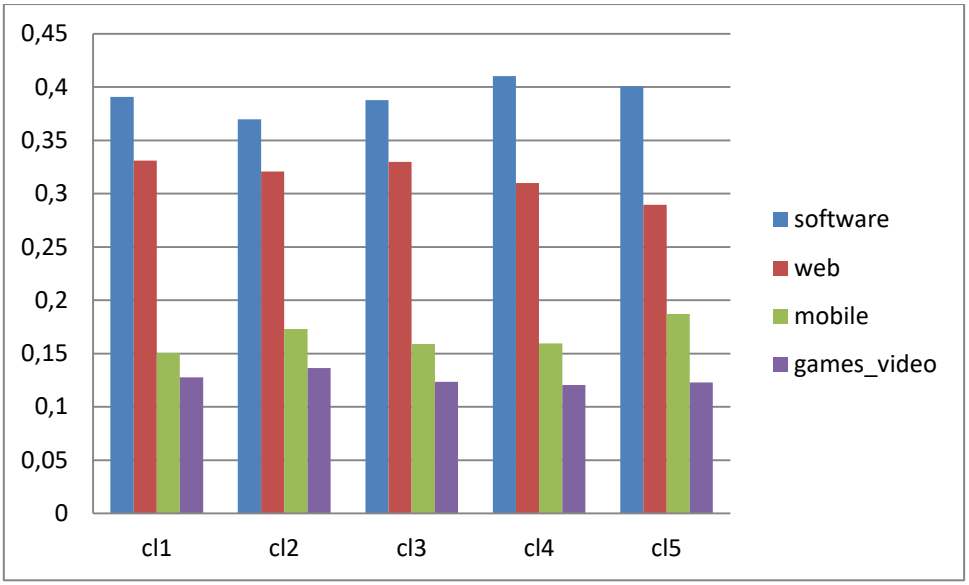
Cluster 4	9.834	11,94
Cluster 5	2.096	2,55
Cluster 6	1.960	2,38
Cluster 7	1.049	1,27
Cluster 8	676	0,82
Cluster 9	648	0,79
Cluster 10	616	0,75
TOTAL	62.817	76,28

Η μέγιστη τιμή αρθρωτότητας είναι $Q = 0.51346$, καθιστώντας τη διαμέριση αποδεκτή. Για την αξιολόγηση των αποτελεσμάτων επιλέχθηκαν οι πέντε (5) πρώτες κατά σειρά μεγέθους κοινότητες, καθώς ορίσθηκε ως κατώφλι οι 2000 κόμβοι ανά ομάδα για την ανάλυση. Ακολουθήθηκαν δυο άξονες για την εξαγωγή συμπερασμάτων, α) διερεύνηση των κόμβων και β) διερεύνηση των κοινοτήτων.

8.2.1 Διερεύνηση των κόμβων

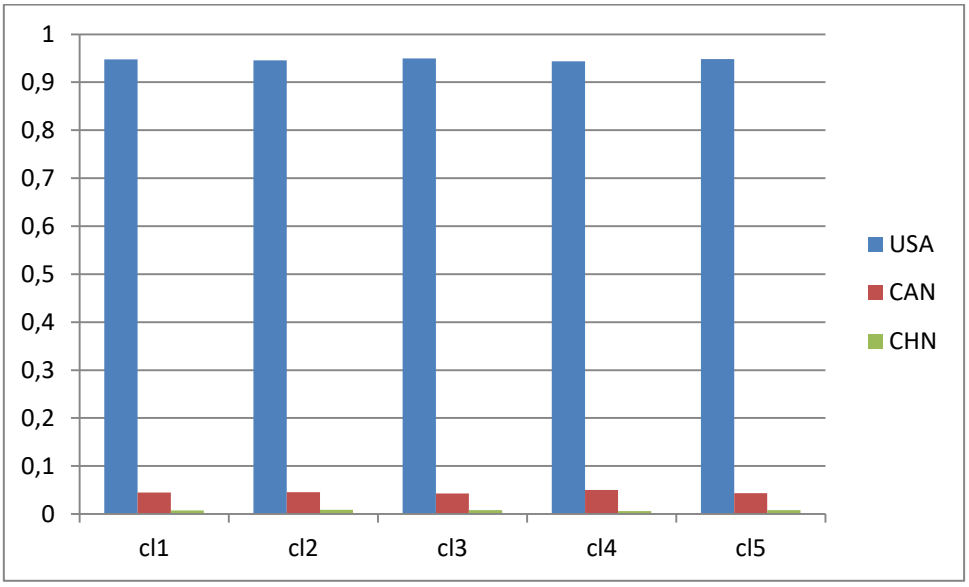
Σε πρώτο στάδιο αναζητήθηκε στους κόμβους εκείνο το χαρακτηριστικό που είναι κοινό εντός των κοινοτήτων. Αρχικά εξετάσθηκε το χαρακτηριστικό «category», το οποίο αναφέρεται στο είδος της δραστηριότητας που ασκεί μια εταιρεία.

Στο γράφημα 8.2 αποτυπώνεται το ποσοστό των κόμβων κάθε κοινότητας που ανήκει στις τέσσερις κατηγορίες με τη μεγαλύτερη συχνότητα.



Γράφημα 8.2 *Ποσοστό κόμβων σε κάθε κατηγορία για 5 κοινότητες.*

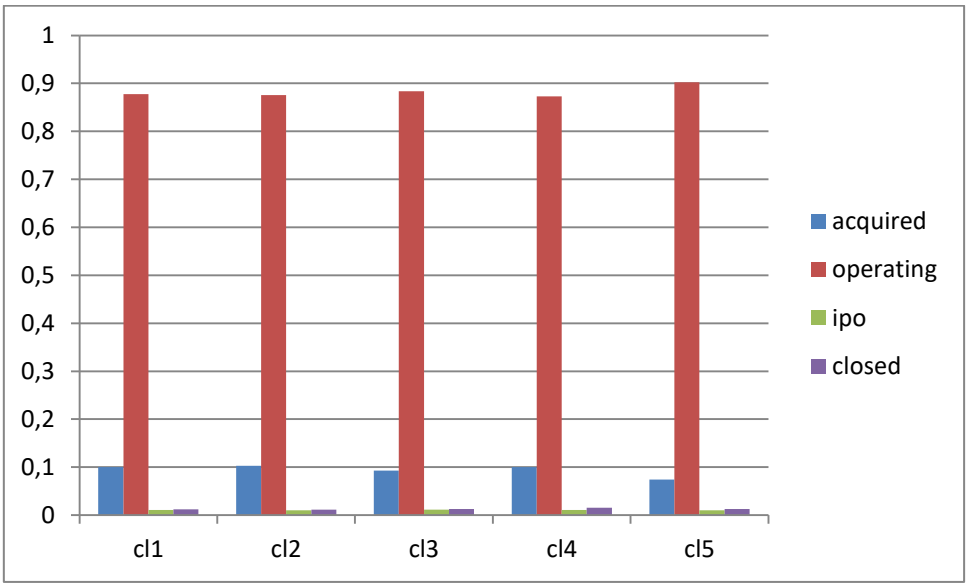
Στη συνέχεια εξετάστηκε το χαρακτηριστικό «country_code», το οποίο αποτυπώνει τη χώρα δραστηριοποίησης της επιχείρησης. Στο γράφημα 8.3 φαίνεται το ποσοστό των εταιρειών που δραστηριοποιούνται σε 3 χώρες για τις 5 μεγαλύτερες ομάδες.



Γράφημα 8.3 *Ποσοστό κόμβων εντός των 5 ομάδων που δραστηριοποιούνται σε ΗΠΑ, Καναδά και Κίνα*

Σε όλες τις ομάδες άνω του 90% των κόμβων δραστηριοποιούνται στις ΗΠΑ, ενώ αντίστοιχα το 5% δραστηριοποιείται στον Καναδά και κάτω του 1% στην Κίνα.

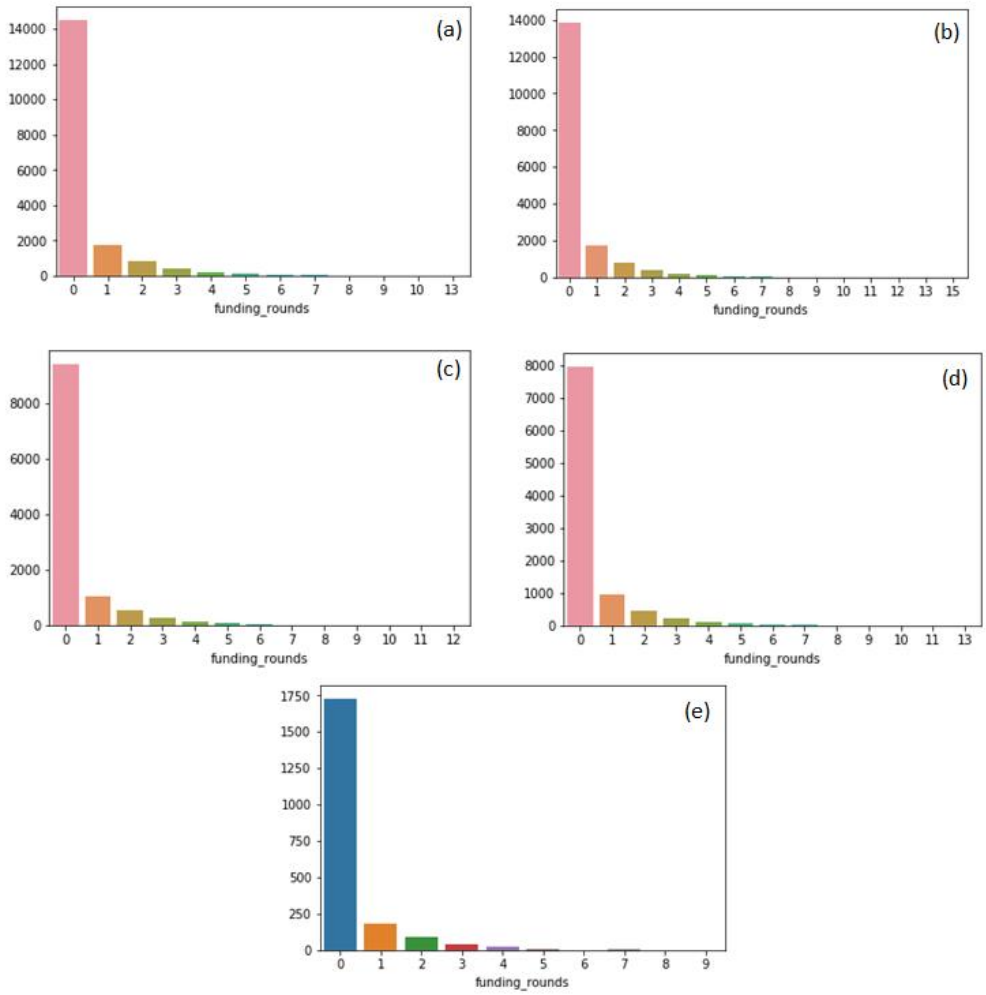
Επόμενο χαρακτηριστικό που εξετάσθηκε είναι το «status», το οποίο δείχνει αν η εταιρεία βρίσκεται στο στάδιο της ανάπτυξης, αν έχει εξαγορασθεί, αν έχει αντλήσει κεφάλαια από την αγορά (Initial Public Offering- IPO) ή αν έχει πτωχεύσει. Στο γράφημα 8.4 αποτυπώνεται το ποσοστό για κάθε μια από τις 4 κατηγορίες, εντός των 5 μεγαλύτερων ομάδων.



Γράφημα 8.4 *Ποσοστό εταιρειών εντός των 5 ομάδων που βρίσκονται στα 4 πιθανά στάδια.*

Παρατηρείται ότι άνω του 85% των εταιρειών κάθε ομάδας βρίσκονται στο στάδιο της ανάπτυξης, οριακά λιγότερο από 10% του δείγματος έχει εξαγορασθεί, ενώ κάτω του 3% των εταιρειών έχει πτωχεύσει ή έχει στραφεί στην αγορά.

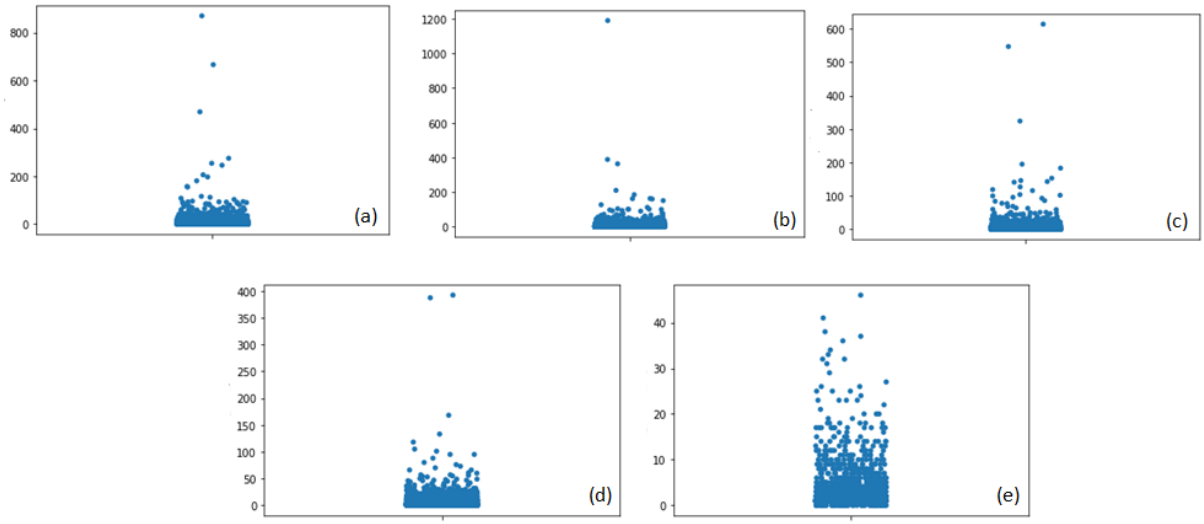
Μια από τις ιδιότητες των κόμβων είναι ο αριθμός των φορών που έχουν αυξήσει τα κεφάλαιά τους από επενδυτές (funding rounds). Στο Γράφημα 8.5 αποτυπώνεται το πλήθος των κόμβων (κάθετος άξονας) ως προς τον αριθμός των χρηματοδοτικών γύρων (οριζόντιος άξονας), για τις πέντε μεγαλύτερες κοινότητες.



Γράφημα 8.5 *a) funding rounds in cl1, b) funding rounds in cl2, c) funding rounds in cl3, d) funding rounds in e) funding rounds in cl5*

Ως προς τα funding rounds των επιχειρήσεων παρατηρείται όμοια κατανομή σε όλες τις ομάδες, αναλογικά με το πλήθος κόμβων της κάθε μιας. Μόνο η ομάδα cl3 διαφοροποιείται, καθώς η μέγιστη τιμή finding rounds είναι 6, ενώ για τις υπόλοιπες η μέγιστη τιμή ισούται με 7.

Το τελευταίο χαρακτηριστικό που εξετάσθηκε εντός των κοινοτήτων είναι ο αριθμός συσχετίσεων. Ο αριθμός αυτός αντιστοιχεί στον πλήθος των ακμών με τις οποίες συνδέεται κάθε εταιρεία στο διμερή γράφο ατόμων- εταιρειών. Το χαρακτηριστικό relationships δηλαδή αντιπροσωπεύει τον αριθμό των ατόμων που έχουν εργασθεί σε κάθε εταιρεία. Στο Γράφημα 8.6 παρουσιάζεται το εύρος των διασυνδέσεων εντός των 5 κοινοτήτων. Φαίνονται κατά σειρά μεγέθους οι 5 πρώτες κοινότητες και στον κάθετο άξονα αποτυπώνεται ο αριθμός διασυνδέσεων.



Γράφημα 8.6: *a) συσχετίσεις κόμβων ομάδας cl1, b) συσχετίσεις κόμβων ομάδας cl2, c) συσχετίσεις κόμβων ομάδας cl3, d) συσχετίσεις κόμβων ομάδας cl4, e) συσχετίσεις κόμβων ομάδας cl5*

Σε όλες τις κοινότητες, η πλειοψηφία των κόμβων εμφανίζουν αριθμό συσχετίσεων με εργαζόμενους μικρότερο του 50. Παράλληλα είναι πολύ μικρός ο αριθμός των κόμβων εντός κάθε κοινότητας που διαφοροποιείται σημαντικά από τους υπόλοιπους ως προς αυτό το χαρακτηριστικό.

Η μοναδική διαφοροποίηση παρατηρείται στο μέγιστο αριθμό συσχετίσεων εντός των ομάδων. Συγκεκριμένα, στις δύο πρώτες ομάδες που είναι και οι μεγαλύτερες φαίνεται να υπάρχουν ορισμένοι κόμβοι με πολύ μεγάλο αριθμό συσχετίσεων (άνω των 800). Ακολούθως, οι δυο επόμενες κατά σειρά μεγέθους κοινότητες περιλαμβάνουν κόμβους με περισσότερες από 300 διασυνδέσεις, ενώ η πέμπτη σε σειρά ομάδα παρουσιάζει μεγαλύτερη ομοιομορφία, με μέγιστο αριθμό διασυνδέσεων να μην ξεπερνά τις 50.

8.2.2 Διερεύνηση των κοινοτήτων

Το δεύτερο στάδιο ανάλυσης επικεντρώθηκε στη δομή των κοινοτήτων που σχηματίσθηκαν από τον αλγόριθμο. Εφόσον παρατηρήθηκε σχετική ομοιομορφία των ομάδων ως προς τα επιμέρους χαρακτηριστικά των κόμβων, με εξαίρεση τον αριθμό

διασυνδέσεων, κρίθηκε σκόπιμο να αναλυθούν οι κοινότητες που προέκυψαν χρησιμοποιώντας μετρικές δικτύου. Ειδικότερα εξετάσθηκε ο μέσος βαθμός κόμβου για κάθε ομάδα, αλλά και η κεντρικότητα κάθε ομάδας ως σύνολο μέσα στο δίκτυο.

Ως προς το μέσο βαθμό κόμβου στις ομάδες, όπως φαίνεται στον πίνακα 8.3 παρατηρείται μεγάλη τιμή στις δυο πρώτες κοινότητες, μεγαλύτερη από τη μέση τιμή βαθμού κόμβου του δικτύου.

Πίνακας 8.3: Μέσος βαθμός κόμβου για 5 κοινότητες και για το δίκτυο

	Average Degree	**Number of nodes**
cl1	14,50654	17.167
cl2	22,403072	17.250
cl3	12,876313	11.521
cl4	8,358145	9.834
cl5	7,307252	2.096
network	12,655925	82300

Το επόμενο χαρακτηριστικό που εξετάσθηκε είναι η κεντρικότητα βαθμού ομάδας, η οποία αποτυπώνει τη συσχέτιση κόμβων που ανήκουν σε διαφορετικές κοινότητες. Μια κοινότητα με μεγάλη τιμή αυτής της μετρικής αποτελείται από κόμβους που έχουν μεγάλο αριθμό συνδέσεων με κόμβους εκτός κοινότητας. Η μετρική αυτή αντιπροσωπεύει την σημαντικότητα μιας κοινότητας εντός του δικτύου.

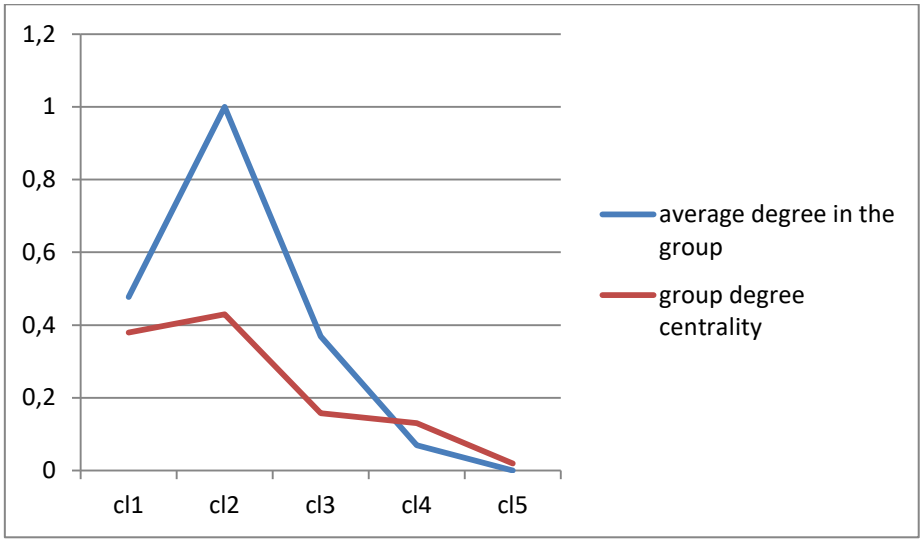
Στον πίνακα 8.4 φαίνεται η κανονικοποιημένη τιμή κεντρικότητας βαθμού ομάδας (group degree centrality) για τις 5 μεγαλύτερες κοινότητες.

Πίνακας 8.4 : Κανονικοποιημένη κεντρικότητα βαθμού των 5 ομάδων

cluster	**group degree centrality**
cl1	0,3796
cl2	0,4297
cl3	0,1580
cl4	0,1307
cl5	0,0196

Βάσει των παραπάνω, η δεύτερη ομάδα (cl2) θεωρείται πιο κεντρική από τις υπόλοιπες, ενώ πολύ κοντινή σε σημαντικότητα φαίνεται η πρώτη ομάδα (cl1).

Στο γράφημα 8.7 αποτυπώνεται η κεντρικότητα των κόμβων εντός των ομάδων σε σύγκριση με την κεντρικότητα των ομάδων στο δίκτυο.



Γράφημα 8.7: *Μέσος βαθμός κόμβων εντός κοινότητας και κεντρικότητα βαθμού κοινότητας*

Η τάση στη μεταβολή του βαθμού τόσο εντός, όσο και μεταξύ των ομάδων είναι όμοια. Η κοινότητα cl2 έχει το μεγαλύτερο μέσο βαθμό κόμβου και είναι η πιο κεντρική κοινότητα στο δίκτυο. Ωστόσο, η κεντρικότητα των ομάδων 4 και 5 είναι χαμηλότερη από το μέσο βαθμό των κόμβων που τις αποτελούν.

Όπως αποδείχθηκε από την εφαρμογή των μεθόδων, ο αλγόριθμος DBSCAN δεν οδηγεί σε αποδοτική ανίχνευση των κοινοτήτων του δικτύου, σε αντίθεση με τη μέθοδο μεγιστοποίηση της αρθρωτότητας, η οποία έδωσε καλύτερα αποτελέσματα.

9

Συμπεράσματα

Στο κεφάλαιο αυτό συνοψίζονται τα αποτελέσματα της εφαρμογής μεθόδων ανάλυσης κοινωνικών δικτύων στο παγκόσμιο δίκτυο νεοφυών επιχειρήσεων και γίνεται αναφορά στα συμπεράσματα που ανέκυψαν. Επιπλέον παρατίθενται προτάσεις για περαιτέρω έρευνα σχετικά με την εφαρμογή περισσότερων μεθόδων ανάλυσης σύνθετων δικτύων για την εξαγωγή συμπερασμάτων.

9.1 Σύνοψη και συμπεράσματα

Κεντρική ιδέα της παρούσας εργασίας είναι η πρόταση μιας μεθοδολογίας καθοδηγούμενης από τα δεδομένα (data driven) για την υποστήριξη της διαδικασίας λήψης επιχειρηματικών αποφάσεων. Περνώντας από το εγωκεντρικό δίκτυο επιχειρήσεων σε ένα πλήρες σύνθετο δίκτυο αλληλεπιδράσεων, προτείνεται η εφαρμογή τεχνικών ανάλυσης κοινωνικών δικτύων για την ανακάλυψη «κρυφών» ιδιοτήτων και προτύπων που συνδέουν τις οντότητες μέσα σε αυτό το δίκτυο.

Για την εφαρμογή της μεθοδολογίας επιλέχθηκε το Παγκόσμιο Δίκτυο Νεοφυών Επιχειρήσεων, καθώς αποτελεί χαρακτηριστικό παράδειγμα περίπτωσης όπου δεν υπάρχουν επαρκή διαχρονικά δεδομένα για την εξαγωγή ασφαλών συμπερασμάτων με παραδοσιακές μεθόδους υποστήριξης αποφάσεων (ανάλυση χρονοσειρών, ανάλυση τάσεων, κ.λπ.).

Το δίκτυο νεοφυών επιχειρήσεων, πέρα από τις άμεσες συσχετίσεις μεταξύ των οντοτήτων, όπως είναι οι εξαγορές, περιλαμβάνει και έμμεσες συνδέσεις, οι οποίες

βασίζονται σε δυο υποθέσεις. Η πρώτη θεωρεί πως η ροή εργαζομένων εντός του δικτύου αντιπροσωπεύει τη ροή τεχνογνωσίας. Η δεύτερη υποστηρίζει ότι εταιρείες που προσελκύουν τους ίδιους επενδυτές είναι πιθανό να έχουν κάποιο κοινό χαρακτηριστικό. Έτσι, μελετώντας τη δομή του δικτύου και ανακαλύπτοντας πιθανά μοτίβα στον τρόπο διασύνδεσης των κόμβων, μπορούν να εξαχθούν συμπεράσματα για τη δύναμη και τη θέση κάθε επιχείρησης ως προς τις υπόλοιπες.

Η συγκεκριμένη εφαρμογή επικεντρώθηκε στην τάση των κόμβων του δικτύου να σχηματίζουν κοινότητες. Ειδικότερα, επιλέχθηκαν αλγόριθμοι που ομαδοποιούν τους κόμβους σε κοινότητες με κριτήρια πυκνότητας (density based clustering). Αρχικά επιλέχθηκε προς εφαρμογή ο αλγόριθμος DBSCAN, καθώς μπορεί να ανακαλύψει κοινότητες σε μεγάλα δίκτυα, όπως το συγκεκριμένο, κατατάσσει τους κόμβους σε κοινότητες ακαθόριστου μεγέθους και επιπλέον εντοπίζει τον θόρυβο. Ο αλγόριθμος εφαρμόσθηκε τροποποιημένος έτσι ώστε να χρησιμοποιεί υπερβολικές αποστάσεις μεταξύ των κόμβων για να τους κατατάξει σε κοινότητες, καθώς η υπερβολική γεωμετρία αποδίδει με πιο εύληπτο τρόπο την απόσταση μεταξύ δυο κόμβων, ενώ επιπλέον ο υπερβολικός χώρος μπορεί να ενσωματώσει περισσότερη πληροφορία από αντίστοιχους ευκλείδειους. Ωστόσο, στην περίπτωση του DBSCAN, το μέγεθος του δικτύου και το κόστος των υπολογισμών κατέστησαν αρκετά χρονοβόρα την ολοκλήρωση του αλγορίθμου, μη επιτρέποντας την διεξαγωγή πολλαπλών πειραμάτων στο σύνολο του δικτύου. Επίσης, η συντριπτική πλειοψηφία των κόμβων του δικτύου ταξινομήθηκαν στην ίδια κοινότητα. Συνεπώς για τη συγκεκριμένη εφαρμογή κρίνεται ότι χρειάζεται ακόμα πιο ενδελεχής μελέτη των παραμέτρων ενσωμάτωσης στον υπερβολικό χώρο αλλά και των παραμέτρων του DBSCAN, έτσι ώστε να εντοπίζει κοινότητες με νόημα. Στη συνέχεια εφαρμόσθηκε στο αρχικό δίκτυο ο αλγόριθμος Greedy Modularity Communities, ο οποίος παρήγαγε μια ικανοποιητική διαμέριση του δικτύου, με ικανοποιητική μέγιστη τιμή αρθρωτότητας και με σημαντικά μικρότερο χρονικό κόστος. Ωστόσο, στην αναζήτηση εντός των κοινοτήτων του χαρακτηριστικού εκείνου που θα μπορούσε να εξηγεί το διαχωρισμό, δεν κατέστη δυνατό να εξαχθεί ασφαλές συμπέρασμα από τη συγκεκριμένη εφαρμογή.

9.2 Προτάσεις για περαιτέρω έρευνα

Όπως έχει ήδη αναφερθεί, η εργασία αυτή επικεντρώθηκε στη μελέτη του δικτύου ως προς την ιδιότητα των κόμβων να δημιουργούν κοινότητες και επιλέχθηκαν αλγόριθμοι που ομαδοποιούν με κριτήρια πυκνότητας. Σε συνέχεια της ανάλυσης προτείνεται η διερεύνηση ιεραρχικών μεθόδων ομαδοποίησης όπως ο αλγόριθμος των Newman- Girvan. Για την ταχύτερη εφαρμογή του αλγορίθμου συνιστάται η πρότερη ενσωμάτωση του δικτύου στον υπερβολικό χώρο.

Βιβλιογραφία

[1] M. Kilkenny και N. Fuller-Love, «Network analysis and business networks,» *Int. J. of Entrepreneurship and Small Business, Vol.21, No.3,* pp. 303-316, 2014.

[2] «www.wikipedia.org,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Startup_company.

[3] M. Bonaventura, V. Ciotti, P. Panzarasa, S. Liverani, L. Lacasa και V. Latora, «Predicting success in the worldwide start-up network,» *Scientific Reports Vol.10, No. 345,* pp. 1-52, 2020.

[4] V. Karyotis, E. Stai και S. Papavassiliou, Evolutionary Dynamics of Complex Communications Networks, CRC Press, 2014.

[5] Chung-Yuan Huang, Chuen-Tsai Sun και Hsun-Cheng Li, «Influence of Local Information on Social Simulations in Small-World Network,» *Journal of Artificial Societies and Social Simulation vol. 8(4),* pp. 1-8, 31 October 2005.

[6] E. N. Gilbert, «Random Graphs,» *Ann. Math. Statist. ,Volume 30, no. 4,* pp. 1141-1144, 1959.

[7] P. Erdos και . A. Renyi, «On random graphs i,» *Publicationes Mathematicae,,* p. 290–297, 19 November 1958.

[8] D. J. Watts και S. Strogatz, «Collective dynamics of ‘small-world’ networks,» *Nature Vol. 393,* pp. 440-442, 1998.

[9] J. D. Medaglia, «Graph Theoretic Analysis of Resting State fMRI,» *Neuroimaging Clin N Am,* p. 593–607, 6 September 2017.

[10] D. Easley και J. Kleinberg, Networks, Crowds, and Markets: Reasoning about a Highly Connected World, Cambridge University Press, 2010.

[11] Albert-László Barabási, «Scale-Free Networks: A Decade and Beyond,» *Science Vol. 325, Issue 5939,* pp. 412-413, 24 July 2009.

[12] A.-L. Barabasi και R. Albert, «Emergence of Scaling in Random Networks,» *Science 286,* pp. 509-512, 21 October 1999.

[13] M. E. J. Newman, «The structure and function of complex networks,» *SIAM*

REVIEW, Vol. 45, No. 2 , pp. 167-256, March 2003.

[14] S. H. Strogatz, «Exploring complex networks,» *NATURE, Vol. 410,* pp. 268-276, 8 March 2001.

[15] «Network-Science,» [Ηλεκτρονικό]. Available: http://www.network-science.org/powerlaw_scalefree_node_degree_distribution.html.

[16] L. C. Freeman, «Centrality in social networks conceptual clarification,» *Social Networks 1,* pp. 215-239, 1979.

[17] L. C. Freeman, «A Set of Measures of Centrality Based on Betweenness,» *Sociometry Vol. 40, No. 1,* pp. 35-41, 1977.

[18] M. Girvan και M. E. J. Newman, «Community structure in social and biological networks,» *PNAS, Vol. 99, No. 12,* pp. 7821-7826, 11 June 2002.

[19] F. D. Malliaros και M. Vazirgiannis, «Clustering and community detection in directed networks: A survey,» *Physics Reports 533* , p. 95–142, 2013.

[20] S. E. Schaeffer, «Graph clustering,» *Computer Science Review 1,* pp. 27-64, 2007.

[21] M. E. J. Newman και M. Girvan, «Finding and evaluating community structure in networks,» *Physical Review, Vol. 69,* pp. 026113-1-026113-15, 2004.

[22] A. Alamsyah , R. Budi και Kuspriyanto, «Community Detection Methods in Social Network Analysis,» *Advanced Science Letters, Vol. 4,* p. 400–407, 2011.

[23] C. Bron και J. Kerbosch, «Algorithm 457: finding all cliques of an undirected graph,» *Communications of the ACM, Volum 16, Issue 9,* p. 1973.

[24] G. Palla, I. Derenyi, I. Farkas και T. Vicsek, «Uncovering the overlapping community structure of complex networks in nature and society,» *arXiv:physics/0506133v1,* 2005.

[25] S. P. Lloyd, «Least squares quantization in PCM,» *IEEE Transactions on Information Theory, Vol 28, No 2,* pp. 129-137, March 1982.

[26] S. Johnson, «Hierarchical clustering schemes,» *Psychometrika, Vol. 32, issue 3,,* pp. 241,254, 1967.

[27] M. Ester, H.-P. Kriegel, J. Sander και X. Xu, «A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,» *KDD'96: Proceedings of the Second International Conference on Knowledge Discovery*

and Data Mining, p. 226–231, 1996.

[28] «towards data science,» [Ηλεκτρονικό]. Available: https://towardsdatascience.com/dbscan-algorithm-complete-guide-and-application-with-python-scikit-learn-d690cbae4c5d.

[29] «www.cs.csi.cuny.edu/,» [Ηλεκτρονικό]. Available: http://www.cs.csi.cuny.edu/~gu/teaching/courses/csc76010/slides/Clustering%20Algorithm%20by%20Vishal.pdf.

[30] A. Clauset, M. Newman και C. Moore, «Finding community structure in very large networks,» *Physical Review Vol. 7, issue 6,* 2004.

[31] V. D. Blondel, J.-L. Guillaume, R. Lambiotte και E. Lefebvre, «Fast unfolding of communities in large networks,» *Journal of Statistical Mechanics Theory and Experiment,* 2008.

[32] H. Cai, V. W. Zheng και K. Chen-Chua, «A Comprehensive Survey of Graph Embedding: Problems, Techniques and Applications,» *IEEE Transactions on Knowledge and Data Engineering, Vol. XX, No. XX,* pp. 1-20, 2018.

[33] X. Zhao, A. Sala, C. Wilson, H. Zheng και B. Y. Zhao, «Orion: Shortest Path Estimation for Large Social Graphs,» *WOSN'10: Proceedings of the 3rd Wonference on Online social networks,* 2010.

[34] S. G. Kobourov και M. Landis, «Morphing Planar Graphs in Spherical Space,» *Journal of Graph Algorithms and Applications, Vol. 12, No. 1,* pp. 113-127, 2007.

[35] X. Zhao, A. Sala, H. Zheng και B. Y. Zhao, «Efficient Shortest Paths on Massive Social Graphs,» *7th International Conference on Collaborative Computing: Networking, Applications and Works (CollaborateCom),* pp. 77-86, 2011.

[36] G. Alanis-Lobato, P. Mier και M. A. Andrade-Navarro, «Efficient embedding of complex networks to hyperbolic space via their Laplacian,» *Scientific Reports, 6:30108,* pp. 1-10, 2016.

[37] G. García-Pérez, A. Allard, M. Á. Serrano και M. Boguñá, «Mercator: uncovering faithful hyperbolic embeddings of complex networks,» *New Journal of Physics, Vol. 21, No. 12,* 2019.

[38] D. Krioukov, F. Papadopoulos, M. Kitsak, A. Vahdat και M. Boguñá, «Hyperbolic geometry of complex networks,» *Physical Review, Vol.82, No. 3,* pp. 036106-1- 036106-18 , 2010.

[39] «pi.math.cornell.edu,» [Ηλεκτρονικό]. Available:

http://pi.math.cornell.edu/~web4520/CG15-0.pdf.

[40] «www.wikipedia.org,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Poincar%C3%A9_disk_model.

[41] B. Wilson και M. Leimeister, «Gradient descent in hyperbolic space,» arXiv:1805.08207v2 [math.OC] , 13 August 2018.

[42] «bjlkeng.github.io,» [Ηλεκτρονικό]. Available: http://bjlkeng.github.io/posts/hyperbolic-geometry-and-poincare-embeddings/.

[43] E. Stai, K. Sotiropoulos, V. Karyotis και S. Papavassiliou, «Hyperbolic embedding for efficient computation of path centralities and adaptive routing in large-scale complex commodity networks,» *IEEE Transactions on Network Science and Engineering 4 (3),* pp. 140-153, 2017.

[44] K. Tsitseklis, M. Krommyda, V. Karyotis, V. Kantere και S. Papavassiliou , «Scalable Community Detection for Complex Data Graphs via Hyperbolic Network Embedding and Graph Databases,» *IEEE Transactions on Network Science and Engineering,* 2020.

[45] «cs.hse.ru,» [Ηλεκτρονικό]. Available: https://cs.hse.ru/data/2015/05/14/1098547089/4._Centrality_Metrics.pdf.

[46] M. Newman, «Fast algorithm for detecting community structure in networks,» *Physical Review Vol. 69, Issue 6,* p. 066133, 2004.