Wikipedia hyperlinks network

Τριφηνόπουλος Χρήστος
ΑΕΜ 9440

- **Εισαγωγή**

Δημιούργησα έναν κώδικα σε python ο οποίος δημιουργεί έναν κατευθυνόμενο γράφο G του οποίου οι κόμβοι είναι τίτλοι άρθρων από το Wikipedia. Κάθε ακμή (u,v) υπάρχει αν το άρθρο με τίτλο u “δείχνει” στο άρθρο v μέσω ενός συνδέσμου στο περιεχόμενο του. Το θέμα που επέλεξα για να τρέξω τον κώδικα μου είναι το *“Gödel's incompleteness theorems”.* Η ανάλυση του δικτύου τόσο μέσω της networkx όσο και μέσω του gephi έδωσε αρκετές ενδιαφέρον πληροφορίες για την φύση των δικτύων που δημιουργούνται από τα hyperlinks της Wikipedia αλλά και για το θέμα που διάλεξα.

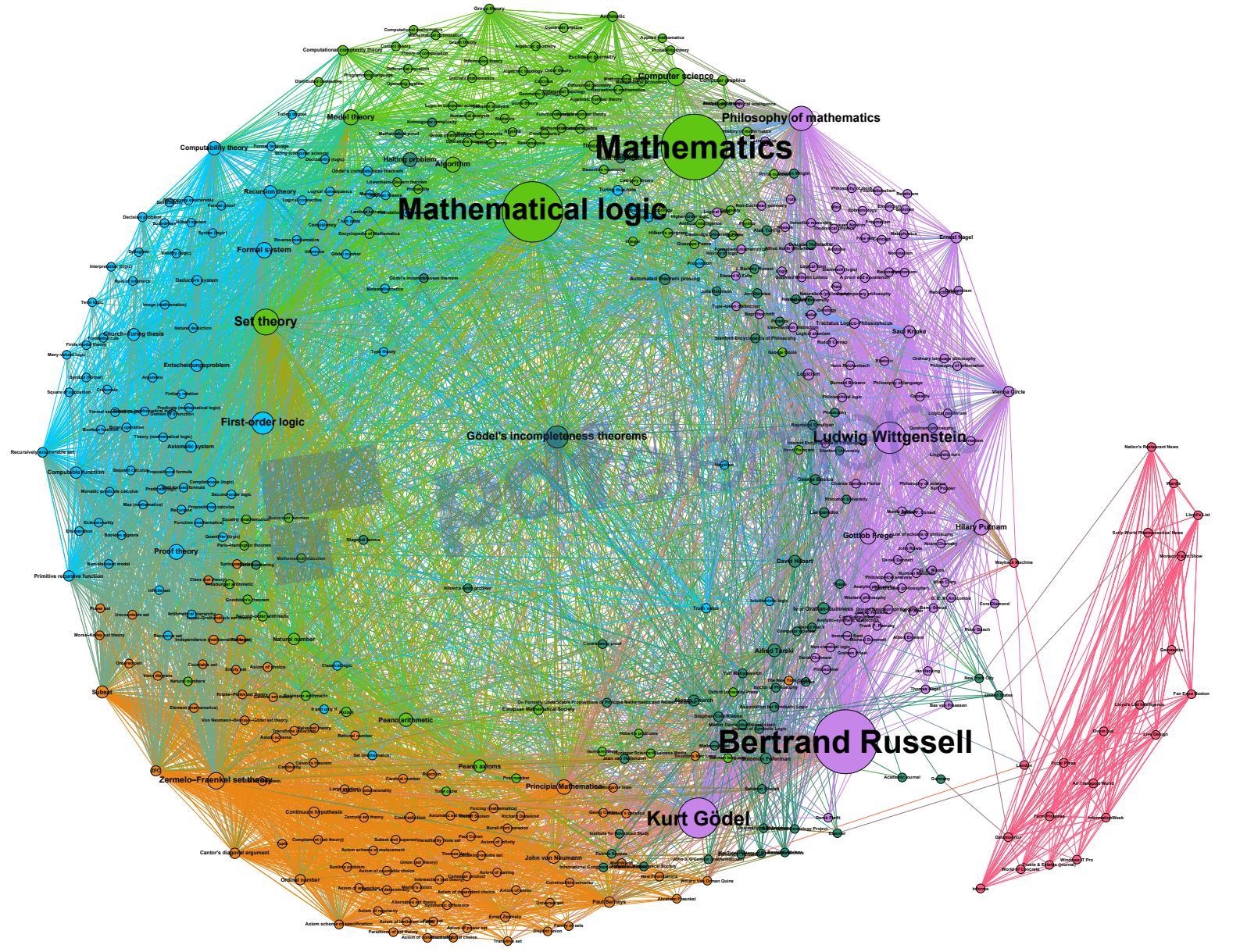
1. Κώδικας Python

Η κύρια συνάρτηση που υλοποίησα είναι η dfs η οποία ψάχνει το δίκτυο της Wikipedia κατά πλάτος χρησιμοποιώντας μια FIFO λίστα (queue). Καθώς ο χρόνος εκτέλεσης αυξάνεται εκθετικά σε σχέση με το βάθος αναζήτησης και η φύση του δικτύου το επιτρέπει επέλεξα να μην τερματίζει όταν ο κώδικας φτάσει σε κάποιο συγκεκριμένο βάθος αλλά όταν βγει από την queue ένας αριθμός άρθρων ο οποίος ορίζεται από το όρισμα “halt” της συνάρτησης. Αν και η χρονική πολυπλοκότητα του κώδικα δεν είναι μεγάλη, αναγκαζόμαστε να ορίσουμε έναν μικρό αριθμό επαναλήψεων καθώς η επανάληψη της συνάρτησης *wikipedia.page(s).links* σε loop δημιουργεί πολλαπλά DNS queries τα οποία καθυστερούν σημαντικά τον κώδικα.
Το άλλο όρισμα της dfs είναι το αρχικό θέμα της αναζήτησης.

Τρέχοντας τον κώδικα για διάφορα inputs ανακάλυψα την ανάγκη για την υλοποίηση ενός φίλτρου (FilterIdentifiers) καθώς τα book identifiers (ISBN κλπ) από την βιβλιογραφία των άρθρων αποτελούσαν σημαντικούς κόμβους του δικτύου ενώ δεν σχετίζονται με το θέμα.

Τέλος έχω χρησιμοποιήσει κάποιες συναρτήσεις της networkx οι οποίες δίνουν πληροφορίες για το δίκτυο που έχει δημιουργήσει ο κώδικας αφού πρώτα το αποθηκεύσει σε ένα αρχείο gexf.

(Για το δικό μου γράφημα έθεσα halt=300)



2. Αρχείο Gephi

Καθώς το πλήθος των κόμβων είναι πολύ μεγάλο επέλεξα να χρησιμοποιήσω φίλτρο για να αποκλείσω κόμβους με indegree μικρότερο του 16. Αυτό το φίλτρο αποδείχθηκε σημαντικό για την δημιουργία ενός γράφου ο οποίος είναι ικανοποιητικός αισθητικά και δεν περιέχει πολλούς κόμβους άσχετους με το αρχικό θέμα.

Μια σημαντική πληροφορία που παίρνουμε (και οπτικά) από το gephi είναι το γεγονός πως το δίκτυο έχει υψηλό average clustering coefficient (0.54). Αυτό επιβεβαιώνει την απλή εμπειρική παρατήρηση ότι τα άρθρα της Wikipedia που προκύπτουν από το ίδιο άρθρο “αλληλοδείχνονται” δημιουργώντας closed triplets.

Για το μέγεθος των κόμβων επέλεξα την ιδιότητα betweenness centrality καθώς είναι ένας πολύ καλός δείκτης για την σημαντικότητα του εκάστοτε κόμβου στο δίκτυο μας.

Το partition του γράφου σε modularity classes χώρισε τους κόμβους με αρκετά μεγάλη ακρίβεια σε ομάδες με βάση το αντικείμενο με το οποίο σχετίζονται τα άρθρα με τα αντίστοιχα ονόματα. (πχ πορτοκαλί - θεωρεία συνόλων, γαλάζιο - μαθηματική λογική, μοβ - φιλοσοφία)

Άφησα στο δίκτυο το άσχετο με το αρχικό θέμα ροζ modularity class κάτω δεξιά καθώς δείχνει το πως ο τρόπος κατασκευής του γράφου είναι ιδιαίτερα αποτελεσματικός στην αυτοματοποιημένη κατηγοριοποίηση εννοιών.

3. Ανάλυση δικτύου

Από το gephi βλέπουμε ότι το δίκτυο έχει μικρή διάμετρο (4) και μικρό average shortest path (1.886), αναμενόμενο αποτέλεσμα καθώς έχουμε περιορίσει την αναζήτηση του αλγόριθμου μας σε μικρό βάθος(<3) με αποτέλεσμα να υπάρχει τουλάχιστον ένα shortest path <7 για κάθε ζευγάρι κόμβων. Με την αφαίρεση των φίλτρων η αύξηση της διαμέτρου είναι μικρή αναλογικά με την αύξηση του αριθμού των κόμβων, γεγονός που μας επιβεβαιώνει ότι οι περισσότεροι κόμβοι που έχουμε φιλτράρει έχουν μικρή σημαντικότητα στο δίκτυο. Η παρατήρηση αυτή ενισχύεται και από το γεγονός οτι με την αφαίρεση των φίλτρων το clustering coefficient μειώνεται σημαντικά (0.134).

Παρατηρώντας την κατανομή των βαθμών στο δίκτυο βλέπουμε πως υπάρχει αρκετά μεγαλύτερη διασπορά στις τιμές του outdegree distribution από τις τιμές του indegree distribution των κόμβων. Δηλαδή ο αριθμός των hyperlinks που περιέχουν τα διάφορα άρθρα ποικίλει αλλά καθώς όπως ανέφερα νωρίτερα τείνουν να “αλληλοδείχνονται” έχουμε μικρότερη διασπορά στον αριθμό των άρθρων που δείχνουν σε αυτά.