KEGG Undirected Graph Dataset

Σε αυτό το πείραμα χρησιμοποιούνται τα δεδομένα από τα KEGG Pathways σε μορφή undirected γράφου. Επιπλέον χρησιμοποιείται μόνο το main component για αποφυγή outliers και bias.

```
Graph with 4810 nodes and 11577 edges
```

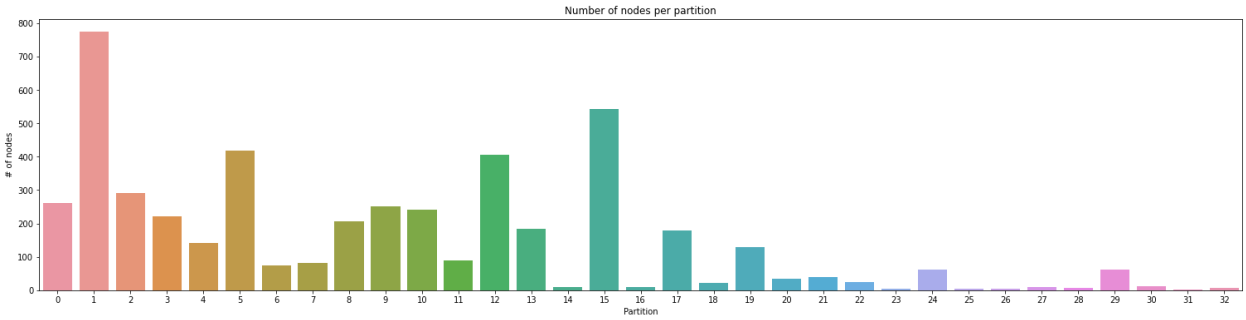
Για την εκπαίδευση του μοντέλου BERT στο task της συσχέτισης γονιδίων και μεταβολητών χρειάζεται ένα train dataset και ένα test dataset. Κύριο μέλημα στη δημιουργία των dataset είναι το μοντέλο να μπορεί να γενικεύσει και σε όρους που μπορεί να μην περιλαμβάνονται στα δεδομένα εκπαίδευσης. Για αυτό, είναι σκόπιμο να γίνει διαχωρισμός του γράφου σε δύο connected γράφους, χωρίς κανέναν κοινό κόμβο μεταξύ τους.
Επιπλέον, αν είναι δυνατή η εκπαίδευση του μοντέλου με μικρότερο train set από ότι test set, τότε αυτό θα επιβεβαίωνε την άντληση πληροφορίας του μοντέλου από την φάση του pretraining στο PubMed.

Louvain partitions

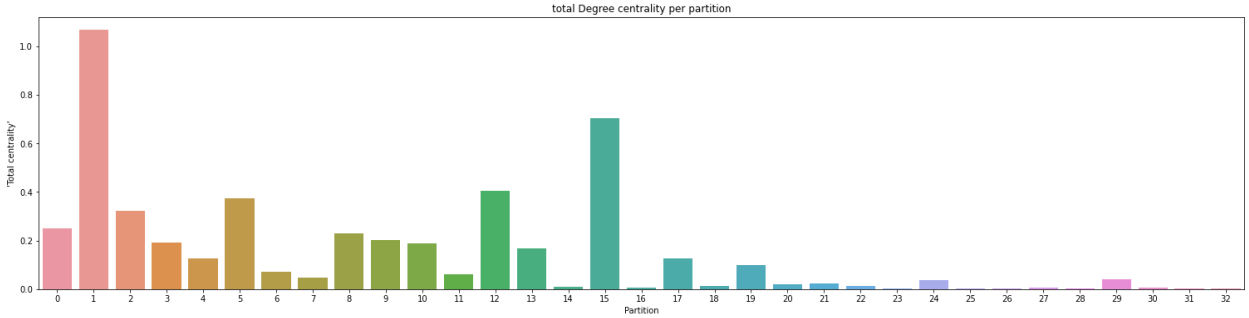
Με βάση τους παραπάνω παράγοντες, χρησιμοποιείται ο αλγόριθμος Louvain για την εύρεση κοινοτήτων στον γράφο. Ιδανικά ο αλγόριθμος θα μπορούσε να χωρίσει τον γράφο κατ' ευθείαν σε δύο και μόνο κοινότητες αλλά στην πράξη αποδείχθηκε ότι η φύση του συγκεκριμένου γράφου δεν το επιτρέπει, παρά την ρύθμιση της παραμέτρου resolution.

Επομένως, ο αλγόριθμος ρυθμίζεται ώστε να δημιουργεί τα ελάχιστα δυνατά partitions. Ο ακριβής αριθμός μπορεί να διαφοροποιείται λίγο σε κάθε εκτέλεση αλλά βρίσκεται κοντά στα 35. Χρησιμοποιείται συγκεκριμένο seed και έτσι δημιουργούνται 32 partitions.

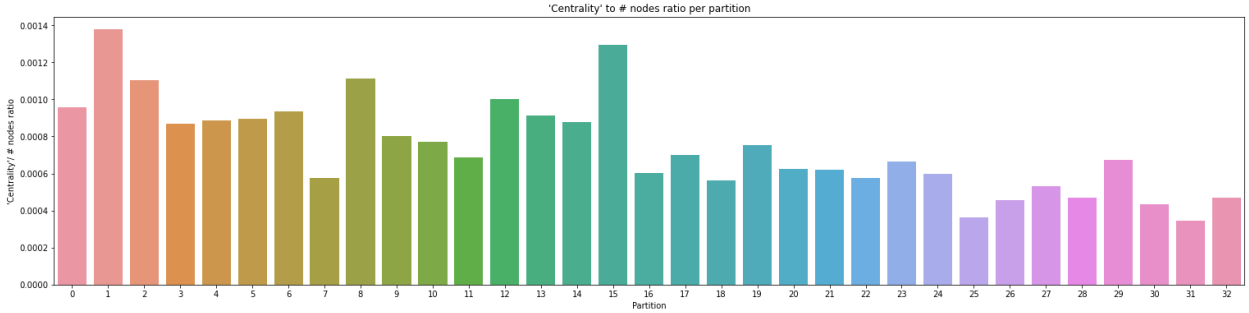
Παρακάτω φαίνεται οι κατανομή των κόμβων ανάμεσα στα 32 partitions με τον αλγόριθμο Louvain:



Μερικά στοιχεία που πιθανόν φανούν χρήσιμα στο επόμενο βήμα είναι το συνολικό centrality του κάθε partition:



Καθώς και ο λόγος του συνολικού centrality προς τον αριθμό κόμβων:



Δημιουργία A/B γράφων

Συνδυάζοντας τα παραπάνω partitions είναι δυνατή η δημιουργία δύο διαφορετικών connected γράφων A και B. Ελλείψει αποδοτικότερης μεθόδου και λόγω πεπερασμένου χρόνου, γίνονται μερικοί τυχαίοι συνδυασμοί των 32 partitions σε 2 ομάδες και κάθε συνδυασμός βαθμολογείται.

Το κριτήριο της βαθμολογίας είναι η ελαχιστοποίηση της παρακάτω σχέσης:

$$\left(\frac{Ngenes_A}{Ncompounds_A} - \frac{Ngenes_B}{Ncompounds_B}\right)^2$$

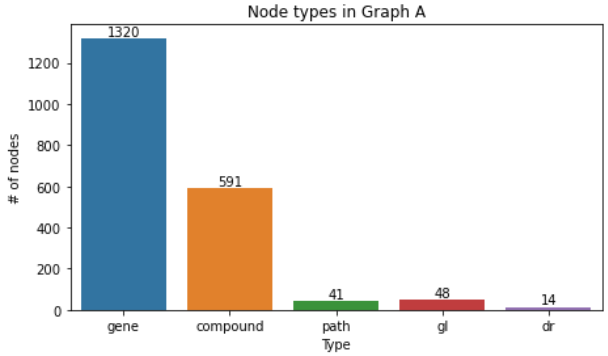
Συνοπτικά το παραπάνω κριτήριο εξασφαλίζει ότι η αναλογία genes προς compounds μεταξύ των δύο γράφων θα είναι παρόμοια. Δηλαδή δεν είναι απαραίτητο να γίνει ισομοιρασμός, αρκεί να τηρούνται οι αναλογίες, ώστε το μοντέλο να εκπαιδευτεί σε αντιπροσωπευτικό αριθμό από κάθε είδος σχέσεων (ECrel, PPrel κλπ).

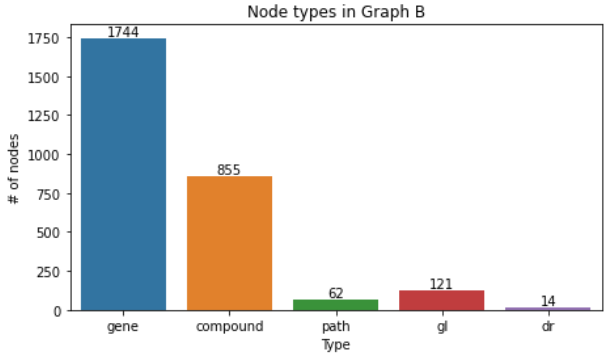
```
Found split(s), best diff=0.03753381446082642 : 100%|██████████████████████████████|
30000/30000 [48:51<00:00, 10.23it/s]
```

Στη συγκεκριμένη εκτέλεση βρέθηκαν 24 συνδυασμοί όπου προκύπτουν connected γράφο. Σε αγκύλες οι κόμβοι του γράφου A:

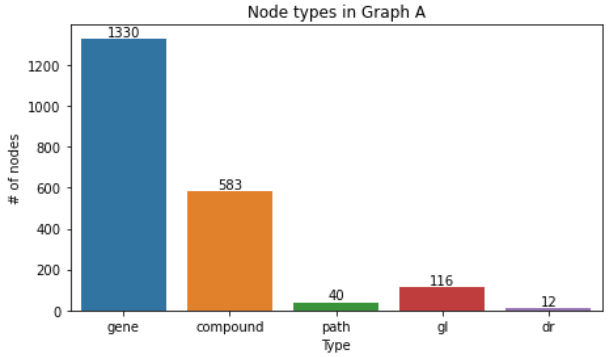
```
Found 37 valid selections
Top 3:
```

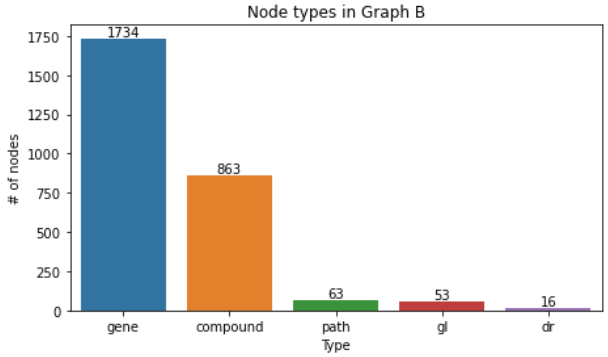
```
Selection #1 (diff= 0.03753381446082642):
[1, 2, 3, 4, 5, 8, 10, 12, 14, 16, 18, 23, 25, 26, 27, 28, 30, 31, 32, 34, 37, 39]
```



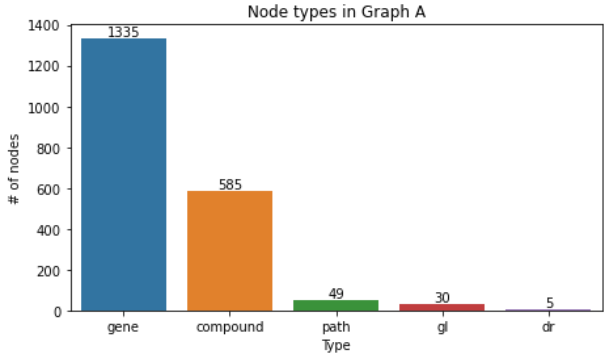


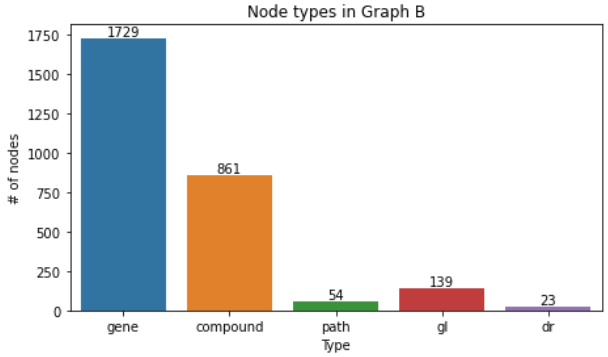

```
Selection #2 (diff= 0.07400228695319476):
[0, 1, 2, 3, 4, 6, 7, 9, 11, 14, 17, 18, 19, 20, 24, 25, 26, 27, 28, 29, 30, 31, 32, 3
7, 38, 39]
```






```
Selection #3 (diff= 0.07503282422057861):
[0, 1, 2, 3, 6, 8, 9, 10, 14, 17, 18, 19, 20, 24, 25, 26, 27, 28, 30, 31, 32, 33, 34,
36, 37, 39]
```





Ο αλγόριθμος αυτόματα επιλέγει την πρώτη από τις 3 περιπτώσεις, αλλά ο χρήστης μπορεί το παρακάμψει και να εισάγει όποιον συνδυασμό επιθυμεί. Παρακάτω επιλέχθηκε ο 1ος συνδυασμός:

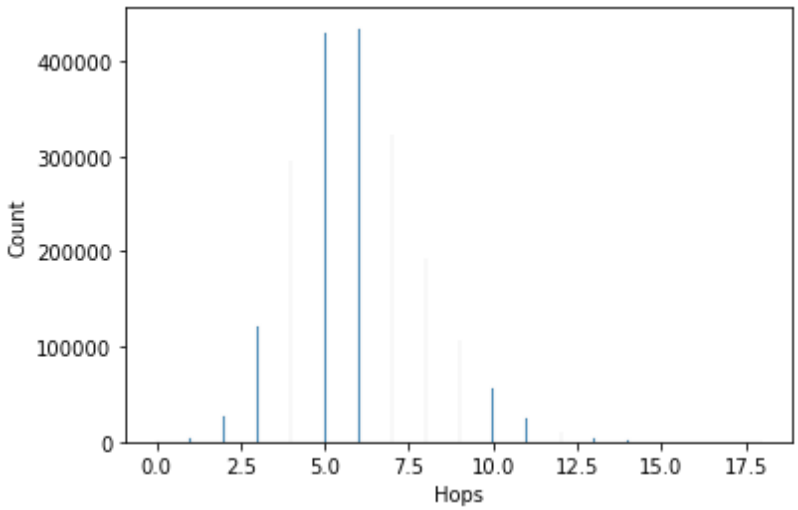
```
Graph A:
Graph with 2014 nodes and 3878 edges
Connected: True
Graph B:
Graph with 2796 nodes and 6507 edges
Connected: True
------------
Common nodes in both Graphs:
------------
```

Δημιουργία Datasets

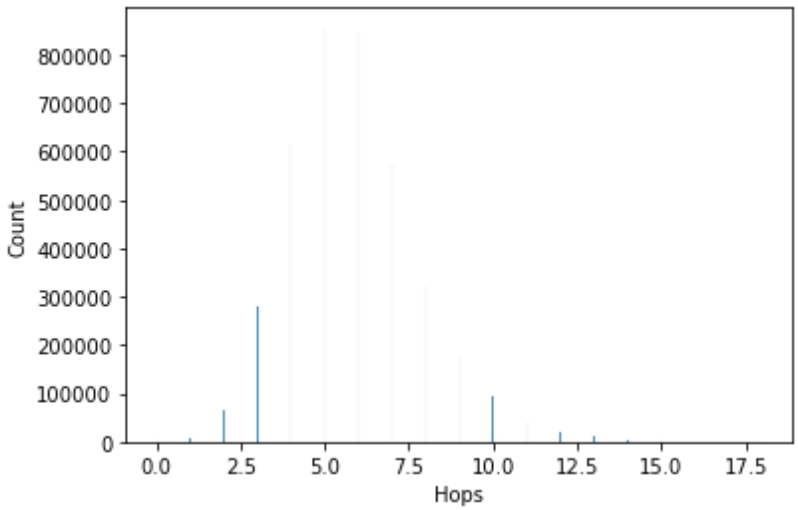
Θα δημιουργηθεί ένα dataset από τον γράφο A και ένα dataset από τον γράφο B.
Σε κάθε γράφο καταγράφονται οι ελάχιστες διαδρομές από κάθε κόμβο του γράφου προς κάθε άλλο κόμβο του ίδιου γράφου.

Έτσι προκύπτουν τα ακόλουθα, για το (Dataset v8):

```
Total Graph A dataset rows:
2029105
```



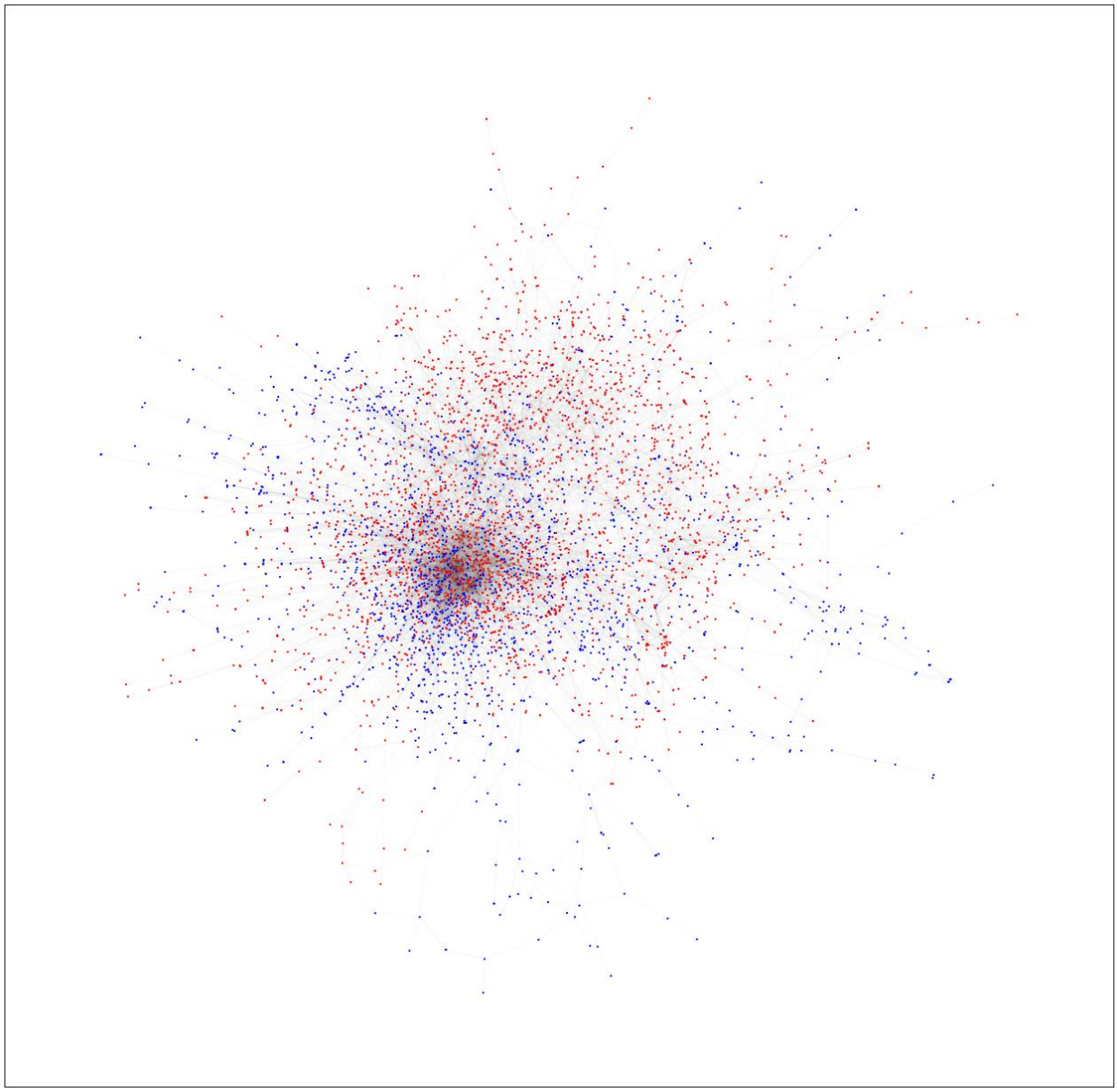

```
Total Graph B dataset rows:
3910206
```



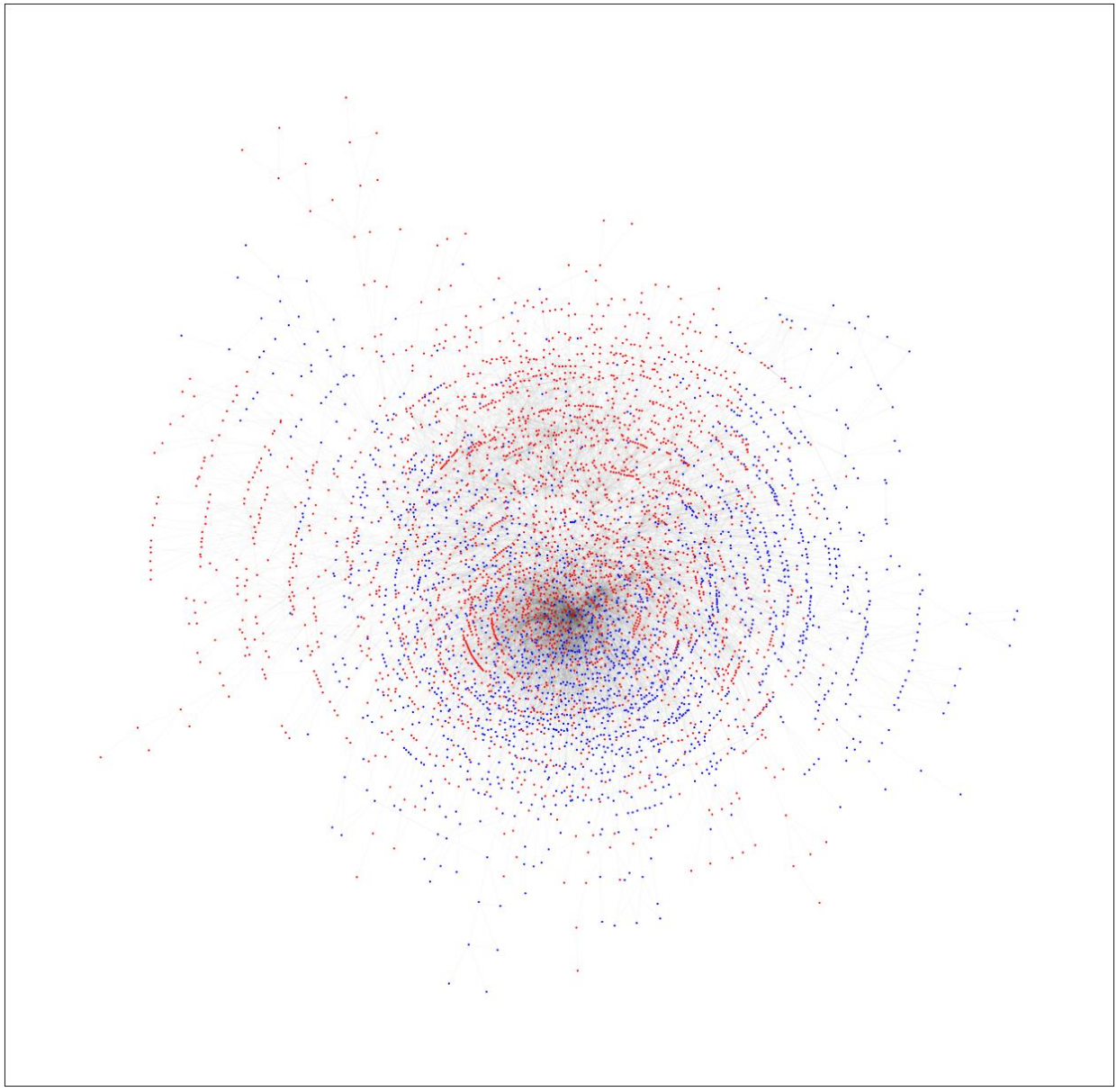
Δεδομένης της υπερ-αντιπροσώπευσης των σχέσεων με ελάχιστη απόσταση 4-7 hops, είναι σκόπιμη η τυχαία δειγματοληψία σε επόμενα βήματα ώστε οι παραπάνω κατανομές να περιέχουν έως 100000 δείγματα για κάθε ελάχιστη απόσταση.
Γίνεται η υπόθεση ότι δεν υπάρχει ουσιαστική καταστροφή πληροφορίας σε μια τέτοια ενέργεια, καθώς στατιστικά είναι πολύ πιθανό ένα ζευγάρι να έχει ελάχιστη απόσταση 4-7 hops, χωρίς όμως να υπάρχει σημαντική συσχέτιση. Διαφορετικά, η πλειονότητα των κόμβων θα είχαν σημαντική συσχέτιση μεταξύ τους.

Οπτικοποίηση κατάτμισης γράφου

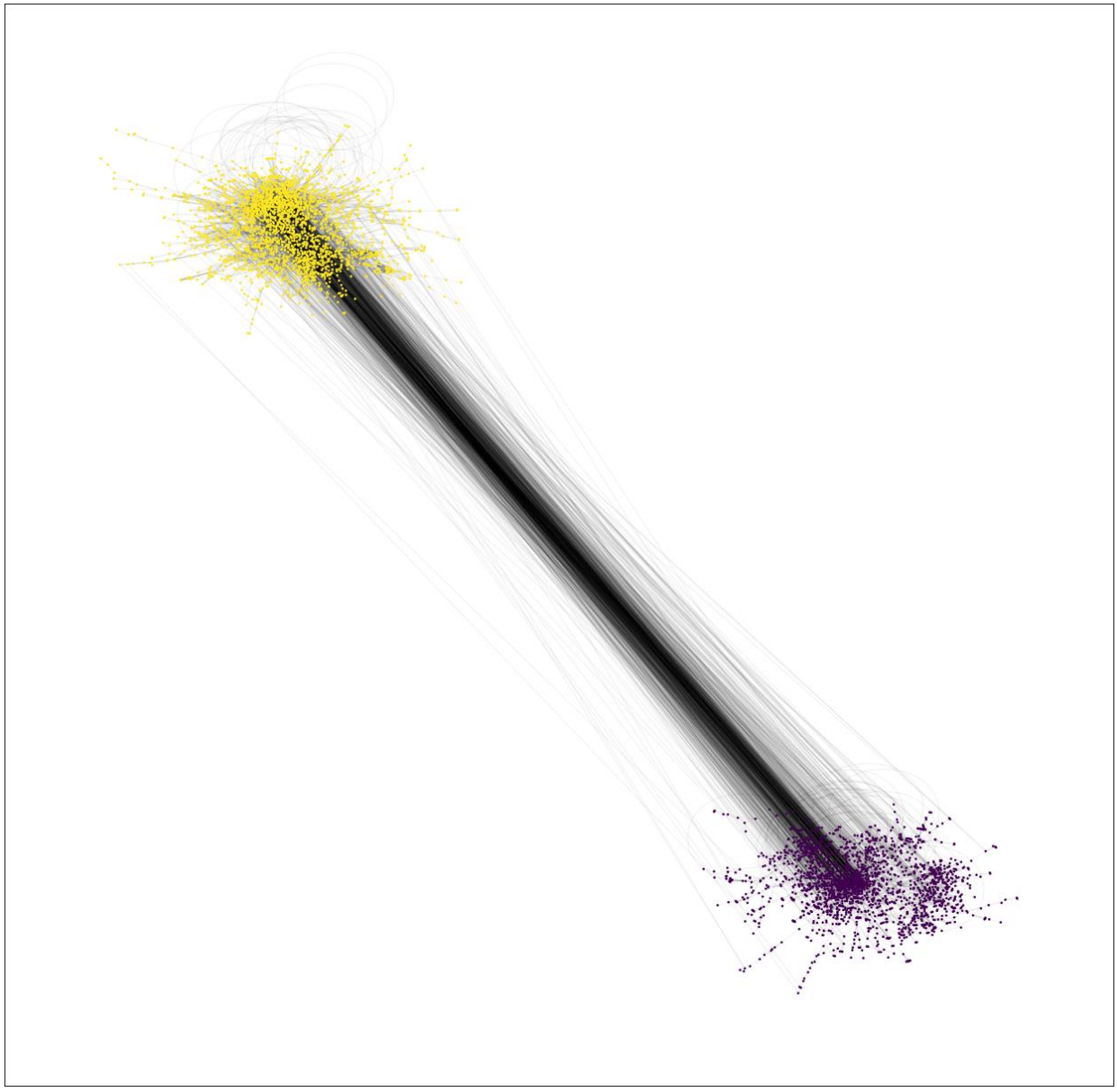
Ενιαίο spring layout:



Ενιαίο Kamada-Kawai layout:



Διαχωρισμένα τμήματα A/B με Spring Layout:



Διαχωρισμένα τμήματα A/B με Kamada Kawai Layout:

Dataset μακρυνών κόμβων

Τέλος, χαρτογραφούνται όλα τα ζευγάρια κόμβων με ελάχιστη απόσταση πάνω από 10 Hops. Αυτά θεωρούνται ως μη συσχετιζόμενα. Σε αυτό το Dataset είναι σημαντικό να υπάρχουν ελάχιστα false positive.

```
Total Distant pairs dataset rows:
292491
```

Ιστόγραμμα: