Graph and Network Analytics

Time-Aware Network Centrality Measures & Link Prediction

Πανεπιστήμειο Πειραιώς

Τμήμα Πληροφορικής

Κυβερνοασφάλεια και Επιστήμη Δεδομένων

Κατεύθυνση: Επιχειρηματική Αναλυτική και Αναλυτική Δεδομένων

Λάμπρος Γανιάς – ΜΠΚΕΔ2203

lampgns96@gmail.com

Νικόλαος Κοκόσης – ΜΠΚΕΔ2211

nikos.kokoshs@yahoo.gr

Ανδρέας Μαρίνης – ΜΠΚΕΔ2219

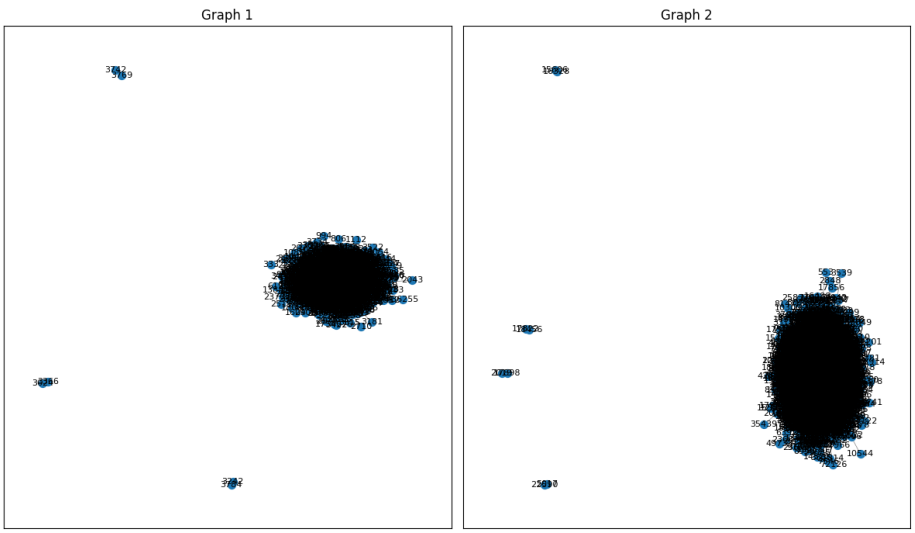
andreasmarinis97@gmail.com

Περίληψη

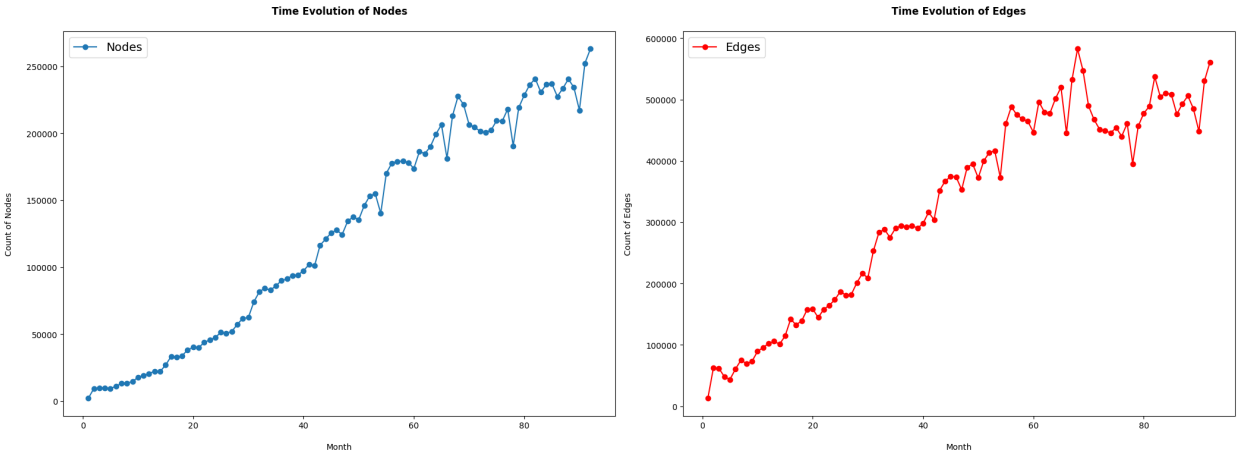
Το πρόβλημα που αναλύεται στην εργασία είναι η ανάλυση του κοινωνικού δικτύου της ιστοσελίδας stackoverflow. Σκοπός της παρακάτω διαδικασίας είναι να φτιάξουμε έναν ταξινομητή ο οποίος θα μπορεί να ταξινομεί αν μια πιθανή ακμή θα γίνει υπαρκτή ή όχι, δηλαδή αν 2 χρήστες θα συνδεθούν ή όχι. Ο ταξινομητής αυτός θα παίρνει ως εισροή μερικές μετρικές της πιθανής ακμής, ενώ θα δίνει ως εκροή την πιθανότητα της ακμής αυτής να γίνει και υπαρκτή. Η διαδικασία προεπεξεργασίας των δεδομένων, με σκοπό την δημιουργία ενός συνόλου δεδομένων το οποίο θα μπορεί να χρησιμοποιηθεί για δυαδική ταξινόμηση, γίνεται αναλύωντας τις συνδέσεις μεταξύ χρηστών που έγιναν σε μια μεγάλη χρονική περίοδο του παρελθόντος. Κατά το πρόβλημα, χωρίζουμε την συνολική χρονική περίοδο σε μικρότερες περιόδους, άρα δημιουργούμε μικρότερους υπο-γράφους. Αντιμετωπίζουμε τον κάθε γράφο ως μη κατευθυνόμενο, ενώ θεωρούμε πως μια ακμή που δημιουργήθηκε σε 2 διαφορετικές χρονικές περιόδους είναι διαφορετική για κάθε μια από τις χρονικές περιόδους αυτές. Θα παρουσιαστεί όλη η διαδικασία προ-επεξεργασίας των δεδομένων, ενώ θα γίνει και η δυαδική ταξινόμηση σε διαφορετικά παραδείγματα από σύνολα δεδομένων που δημιουργήσαμε με διαφορετικές προσεγγίσεις.

1. Διαχωρισμός και Ανάλυση Υπο-Γράφων

Το dataset που μας δόθηκε είχε καταγεγραμμένες συνδέσεις μεταξύ χρηστών για μια περίοδο περίπου 2774 ημερών. Επιλέξαμε, να διαχρωρίσουμε αυτόν τον μεγάλο γράφο σε τόσους μικρότερους γράφους έτσι ώστε κάθε ένας να έχει περίοδο περίπου **30 ημερών / 1μήνα.** Ο λόγος που επιλέξαμε αυτήν την περίοδο είναι ότι θέλαμε μια αρκετά μεγάλη περίοδο η οποία θα επέτρεπε στον κάθε γράφο να έχει αποτυπωμένές όλες τις ιδιότητές του συνολικού δικτύου, ενώ παράλληλα θα έπρεπε ο κάθε ένας γράφος να μην είναι πολύ μεγάλος, για να μην αντιμετωπίσουμε πρόβλημα υπολογιστικά. Τελικά, χωρίζοντας το δίκτυο μας σε "μηνιαίους" γράφους, καταφέραμε να πάρουμε **93 διαδοχικούς γράφους**.Έχοντας χωρίσει το dataset σε 93 διαδοχικούς μικρότερους γράφους, το πρώτο που δοκιμάσαμε είναι να εμφανίσουμε με την βοήθεια της networkX τους πρώτους 2 γράφους:



Βλέπουμε, πως τους πρώτους 2 μήνες του δικτύου μας δεν υπάρχει ιδιαίτερη διαφορά στα δομικά χαρακτηριστικά του γράφου κοιτάζοντας τους με γυμνό μάτι. Αυτό που μπορούμε να παρατηρήσουμε είναι το γεγονός ότι υπάρχει ένας κυρίαρχος συνεκτικός γράφος και κάποιες δυάδες κόμβων που βρίσκονται εκτός αυτού. Στην συνέχεια, προσπαθήσαμε να δούμε πως συμπεριφέρονται οι αριθμοί των κόμβων και των ακμών κατά το πέρασμα του χρόνου:



Παρατηρούμε πως όσο περνάνε οι μήνες, οι χρήστες γίνονται όλο και περισσότεροι ενώ κάτι αντίστοιχο συμβαίνει και για τις συνδέσεις μεταξύ των χρηστών. Βλέπωντας πως ο πληθυσμός μεγαλώνει όσο περνάει ο χρόνος, θα ελέγξουμε τις κατανομές κάποιων μέτρων κεντρικότητας για

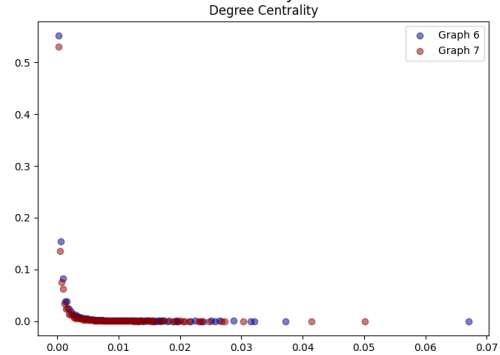
τους κόμβους των γράφων μας με σκοπό να ανακαλύψουμε την δομή του δικτύου μας. Τα μέτρα κεντρικότητας που υπολογίσαμε είναι τα εξής:

1. **Degree Centrality:**
   - Είναι η απλούστερη μέτρηση κεντρικότητας.
   - Υπολογίζει τον αριθμό των ακμών που συνδέονται με έναν κόμβο (τον βαθμό του).
   - Οι κόμβοι με υψηλότερους βαθμούς θεωρούνται πιο κεντρικοί όσον αφορά τις συνδέσεις.
2. **Closeness Centrality:**
   - Μετρά πόσο κοντά βρίσκεται ένας κόμβος σε όλους τους άλλους κόμβους στο δίκτυο.
   - Υπολογίζει τον μέσο συντομότερο μονοπάτι από έναν κόμβο σε όλους τους άλλους κόμβους.
   - Οι κόμβοι με υψηλό closeness centrality είναι πιο προσβάσιμοι και κεντρικοί.
3. **Betwenness Centrality:**
   - Μετρά τον βαθμό όπου ένας κόμβος βρίσκεται στα συντομότερα μονοπάτια μεταξύ άλλων κόμβων.
   - Αναγνωρίζει τους κόμβους που λειτουργούν ως γέφυρες, συνδέοντας διάφορα τμήματα του δικτύου.
   - Οι κόμβοι με υψηλό betwenness centrality συχνά έχουν έλεγχο στη ροή της πληροφορίας.
4. **Eigenvector Centrality:**
   - Λαμβάνει υπόψη την κεντρικότητα των γειτόνων ενός κόμβου.
   - Αντανακλά την ιδέα ότι η σύνδεση με δημοφιλείς κόμβους καθιστά έναν κόμβο πιο δημοφιλή.
5. **Katz Centrality:**
   - Παρόμοια με την κεντρικότητα ιδιοτιμής, αλλά επιτρέπει παραμετροποιημένη επίδραση από τους γείτονες.

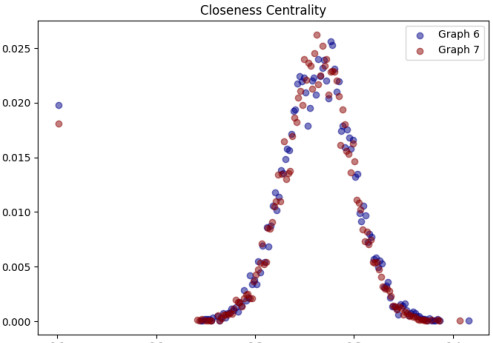
Έχοντας πλέον ορίσει τα μέτρα κεντρικότητας, θα υπολογίσουμε τις κατανομές τους στους διαδοχικούς γράφους 6 & 7 σε κανονική κλίμακα αλλά και σε λογαριθμική (log-log scale). Τα αποτελέσματα παρουσιάζονται παρακάτω:

Normal Scale Probability Distributions

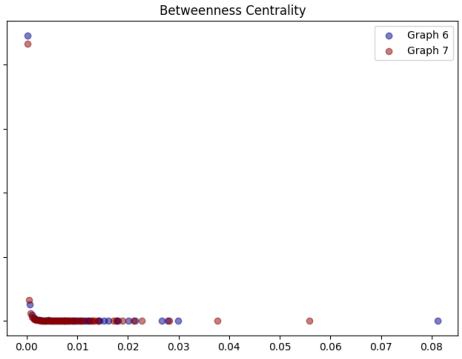
Degree Centrality



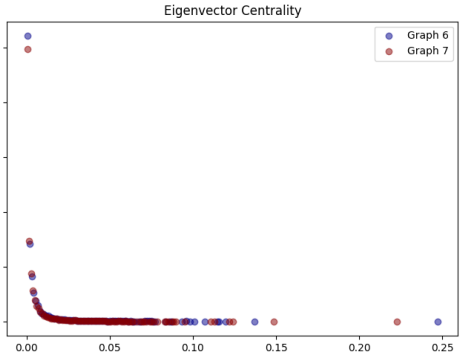
Closeness Centrality



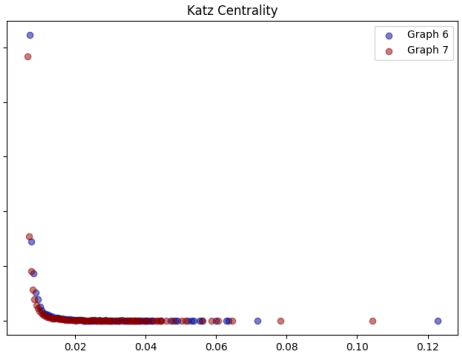
Betweenness Centrality



Eigenvector Centrality

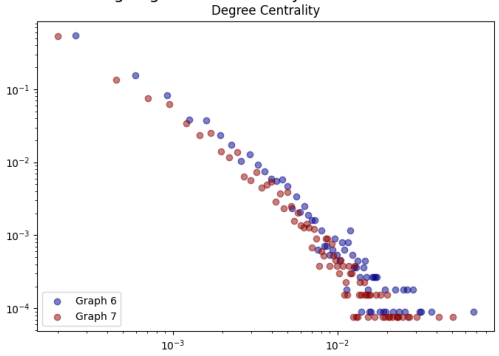


Katz Centrality

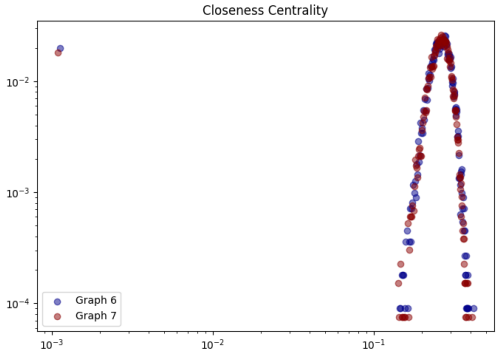


Log-Log Scale Probability Distributions

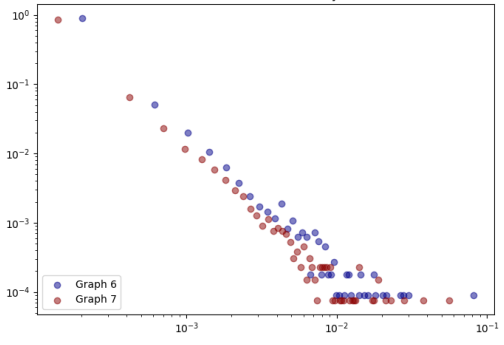
Degree Centrality



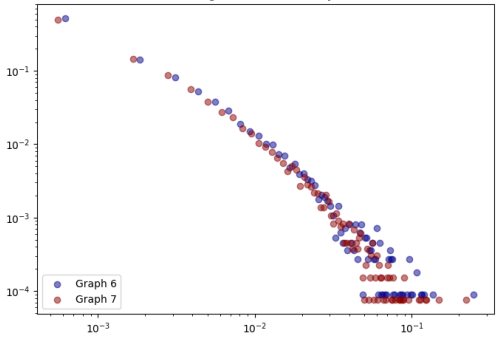
Closeness Centrality



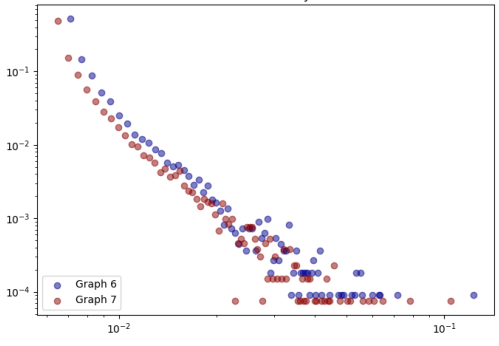
Betweenness Centrality



Eigenvector Centrality



Katz Centrality



Στην **κανονική κλίμακα** παρατηρούμε πως όλα τα μέτρα κεντρικότητας εκτός του closeness centrality ακολουθούν **Power-Law κατανομή**, όπου όσο μικρότερη ειναι η τιμή κεντρικότητας τόσο περισσότεροι κόμβοι εμφανίζονται με την τιμή αυτή.

Στην **log-log κλίμακα,** παρατηρούμε πως ειδικά στο **degree centrality** εμφανίζεται μια μικρή καμπύλη που υποδεικνύει πως ο ρυθμός μείωσης της πιθανότητας αυξάνεται όσο πηγαίνουμε σε μεγαλύτερα μέτρα κεντρικότητας.

Συγκρίνοντας τους 2 διαδοχικούς γράφους μπορούμε να παρατηρήσουμε (ευκολότερα μέσω του log-log scale distribution plot) πως υπάρχει μια μετακίνηση των κατανομών, και ειδικότερα του betwenness centrality, προς χαμηλότερα μέτρα κεντρικότητας. Για να επιβεβαιώσουμε αυτό που βλέπουμε με γυμνό μάτι, θα χρησιμοποιηθεί η μέθοδος KL Divergence για να δούμε πόσο άλλαξε η κάθε κατανομή στον χρόνο. Η KL Divergence απόσταση για ένα διακριτό διάστημα X ορίζεται από τον παρακάτω τύπο:

$$\int_{-\infty}^{\infty} \{\log \frac{P(x)}{Q(x)}\} P(x)dx$$

Όπου Q(x) είναι η πιθανότητα του διαστήματος x να συμβεί στον 1ο από τους 2 διαδοχικούς γράφους, ενώ P(x) να συμβεί στον 2ο. Για να πάρουμε την συνολική απόσταση όλης της κατανομής, αθροίζουμε κάθε KL-DIV του κάθε διακριτού διαστήματος. Πήραμε τα εξής αποτελέσματα:

```
The KL - Divergence for the Graphs 6 & 7 are:
------------------------------------------------------
Degree Centrality: 1.4247068218102161
Closeness Centrality: 0.1470924261622017
Betweeness Centrality: 4.702912758129193
Eigenvector Centrality: 2.0902969195289405
Katz Centrality: 0.1020893323729962
```

Παρατηρούμε πως όντως η μεγαλύτερη συρρίκνωση κατανομής γίνεται από το betwenness centrality ενώ σημαντικές συρρικνώσεις προς τις μικρότερες τιμές έχουν επίσης degree & eigenvector centrality. Το αποτέλεσμα αυτό μας οδηγεί στα εξής συμπεράσματα για το δίκτυο μας:

1. Εφόσον το **degree centrality** συρρικνώνεται σε μικρότερες τιμές κεντρικότητας κατά το πέρασμα του χρόνου καταλαβαίνουμε πως οι νέοι κόμβοι που εισέρχονται στο δίκτυο

προτιμούν να συνδεθούν με τους πιο διάσημους κόμβους κι έτσι εμφανίζεται το φαινόμενο πως “οι πλούσιοι γίνονται πλουσιότεροι και οι φτωχοί παραμένουν φτωχοί”.

2. Αυτή η μεγάλη μεταβολή στο **betwenness centrality** μας δείχνει πως ολοένα και λιγότεροι κόμβοι αποτελούν σημαντικές “γέφυρες” που συνδέουν όλο το δίκτυο μεταξύ τους
3. Η μεταβολή στο **eigenvector centrality** μας δείχνει πως με την πάροδο του χρόνου οι περισσότεροι χρήστες τείνουν να έχουν “λιγότερο δημοφιλείς φίλους”, κάτι που επηρεάζεται από το φαινόμενο πως οι νέοι κόμβοι προτιμούν να συνδεθούν με δημοφιλέστερους κόμβους.

Έχοντας εξάγει τα παραπάνω συμπεράσματα από τις κατανομές μέτρων κεντρικότητας μπορούμε με ασφάλεια να κάνουμε την υπόθεση πως βρισκόμαστε σε ένα **scale-free δίκτυο** όπου:

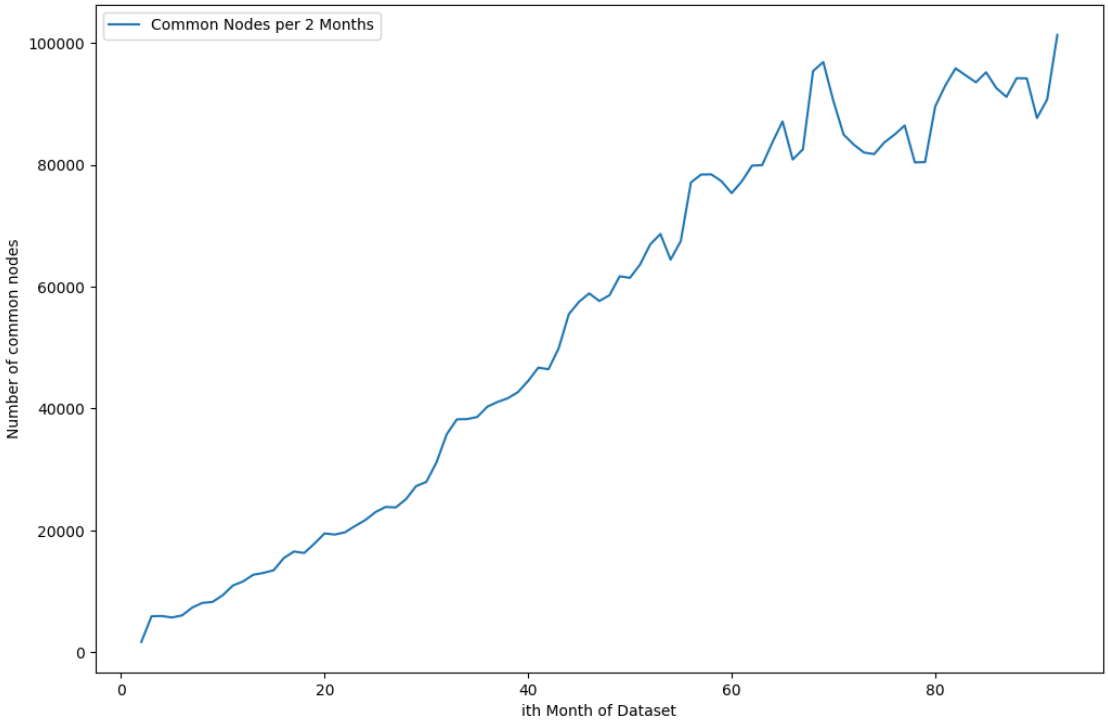
1. Οι κόμβοι αυξάνονται κατά το πέρας του χρόνου
2. Οι νέοι κόμβοι που εισέρχονται στο δίκτυο προτιμούν να συνδεθούν με περισσότερο διάσημους κόμβους.

Πιο συγκεκριμένα, το δίκτυο μας είναι ένα **truncated scale-free network,** λόγω της καμπύλης αυτής που εμφανίζεται στην log-log scale κατανομή του degree centrality. Η καμπύλη αυτή μπορεί να οφείλεται σε:

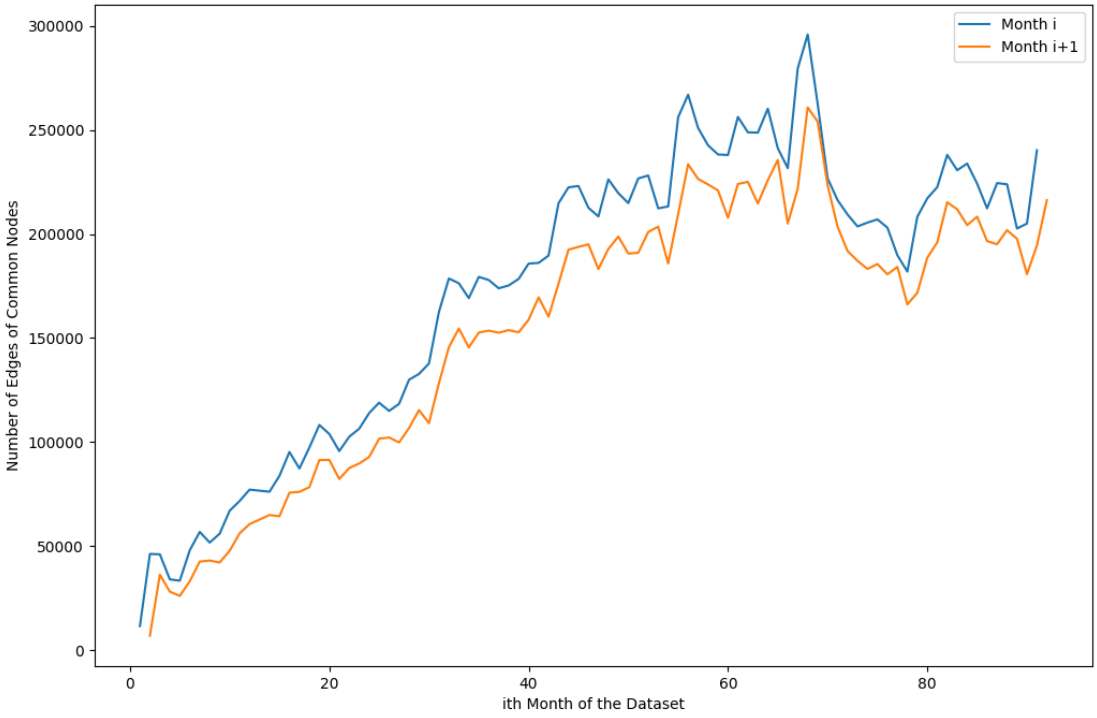
1. **Age Effect:** Οι παλιοί κόμβοι σταματάνε να κάνουν συνδέσεις κατά την πάροδο του χρόνου
2. **Cost Effect:** Το να συντηρεί συνδέσεις ένας κόμβος έχει και κάποια κόστη κι έτσι είναι αδύνατον ένας κόμβος να έχει άπειρες συνδέσεις.

2. Κοινοί κόμβοι, ακμές και ανάλυση

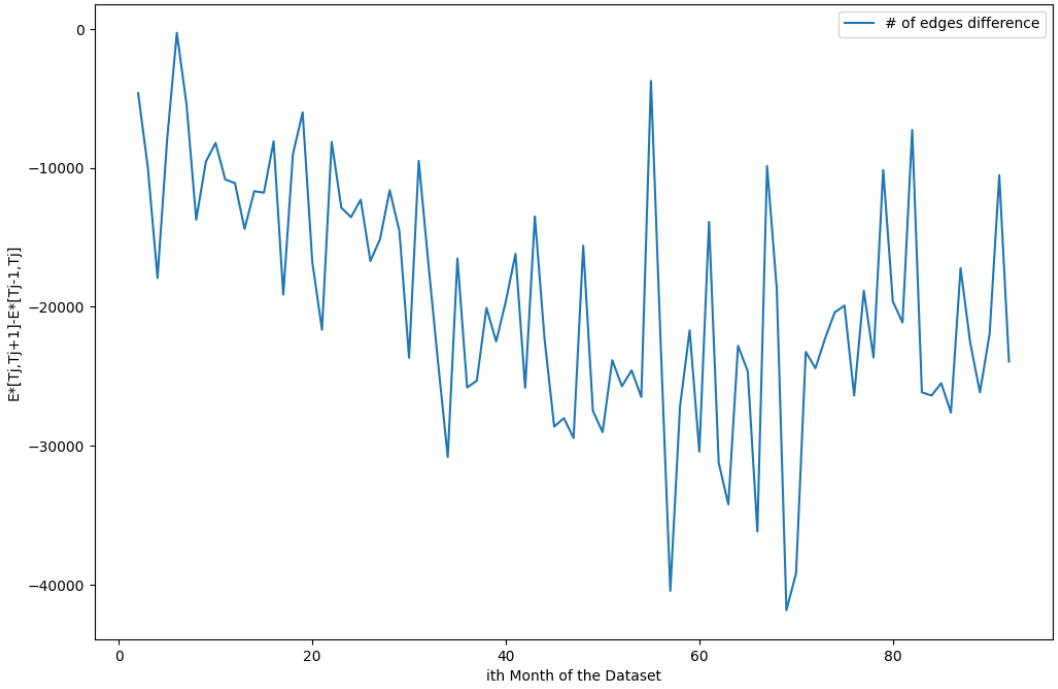
Έχοντας υποθέσει λοιπόν πως βρισκόμαστε σε ένα truncated scale free δίκτυο, σκοπός μας είναι να φτιάξουμε ένα σύνολο δεδομένων με ακμές υπαρκτές και μη, έτσι ώστε να μπορέσουμε να αντιμετωπίσουμε το πρόβλημα της πρόβλεψης σύνδεσης 2 χρηστών σαν πρόβλημα δυαδικής ταξινόμησης. Έτσι, ανά 2 διαδοχικούς γράφους θα κρατάμε μόνο τους κοινούς κόμβους V* που συμμετέχουν και στους 2 γράφους, από τους οποίους κρατάμε τις υπαρκτές ακμές που συνέβησαν στον 1ο γράφο **E*prev σαν δεδομένα εκπαίδευσης** και αυτές που συνέβησαν στον 2ο γράφο **E*next σαν δεδομένα ελέγχου**. Ας δούμε πως συμπεριφέρονται οι ποσότητες των κόμβων και των ακμών αυτών κατά το πέρας του χρόνου:



Οι κοινοί κόμβοι αυξάνονται κατά το πέρας του χρόνου, κάτι αναμενόμενο αφού γενικώς είχαμε δει πως οι κόμβοι αυξάνονται όσο περνάνε οι μήνες.



Όσον αναφορά τον αριθμό των κοινών ακμών βλέπουμε ότι και αυτές κατά το πέρας των μηνών αυξάνονται. Μια ακόμα παρατήρηση που προκύπτει είναι ότι σχεδόν πάντα οι ακμές του E*next είναι λιγότερες από αυτές του E*prev. Αυτό μας διευκολύνει να δώσουμε σαν δεδομένα εκπαίδευσης αυτά του E*prev έτσι ώστε ο ταξινομητής να έχει περισσότερα δεδομένα για να εκπαιδευτεί. Ας δούμε και πως συμπεριφέρεται και η διαφορά των 2 σε ένα γράφημα:



Βλέπουμε ότι όχι μόνο πάντα η διαφορά αυτή είναι αρνητική, αλλά και κατά το πέρασμα του χρόνου μεγαλώνει όλο και περισσότερο ως και περίπου τον 70ο μήνα.

3. Δημιουργία Dataset, Εκπαίδευση & Πρόβλεψη

Σκοπός της άσκησης είναι να δημιουργήσουμε ένα dataset το οποίο θα εμπεριέχει πιθανές ακμές από τις οποίες θα πάρουμε τις εξής μετρικές ομοιότητας:

1. **Graph Distance**: Απόσταση των κόμβων
2. **Common Neighbors**: Αριθμός κοινών γειτόνων
3. **Jaccard Coefficient**: Αριθμός κοινών γειτόνων, κανονικοποιημένος με τον αριθμό της ένωσης των γειτόνων των κόμβων.

4. **Adamic Adar**: Μέτρο ομοιότητας με βάση τους κοινούς γείτονες, δίνοντας περισσότερο έμφαση σε γείτονες οι οποίοι είναι σπανιότεροι.
5. **Preferential Attchment**: Το γινόμενο του αριθμού των γειτόνων των 2 κόμβων.

Επίσης, πέρα από τις μετρικές θα έχουμε και ένα label για κάθε πιθανή ακμή όπου:

- **label = 1,** η πιθανή ακμή ήταν και πραγματική
- **label = 0,** η πιθανή ακμή δεν πραγματοιήθηκε

Όμως ξέρουμε ποιες είναι οι πραγματικές ακμές, δηλαδή αυτές με label = 1. Είναι οι E*prev για το training dataset & οι E*next για το test dataset. Αυτό που μένει να κάνουμε είναι να δημιουργήσουμε εκείνες τις ακμές που δεν πραγματοποιήθηκαν, δηλαδή αυτές που θα έχουν label = 0. Οι ακμές αυτές είναι οι (u,v) για κάποιους κόμβους u & v που βρίσκονται στο σύνολο V*xV*, εκτός από αυτές που πραγματικά συνέβησαν είτε στο E*prev είτε στο E*next και το σύνολο αυτών το συμβολίζουμε ως **Enon**. Όμως, επειδή το Enon είναι πολύ μεγαλύτερο από τα άλλα 2 σύνολα, πρέπει να βρούμε κάποιον τρόπο να αντλήσουμε ένα δείγμα (negative sampling) **Enon_prev** ίσου μεγέθους με το E*prev, κι ένα **Enon_next** ίσου μεγέθους με το E*next αντίστοιχα, έτσι ώστε να έχουμε ένα ισορροπημένο σύνολο δεδομένων με υπαρκτές και μη υπαρκτές ακμές. Για την δειγματοληψία των μη υπαρκτών ακμών ακολουθήσαμε τρεις διαφορετικές στρατηγικές και συγκρίναμε τα αποτελέσματα που μας απέφεραν. Οι 3 στρατηγικές που ακολουθήσαμε ήταν οι εξής:

1. Λιγότερο Δημοφιλής Κόμβος - Γείτονας Δημοφιλέστερου Κόμβου

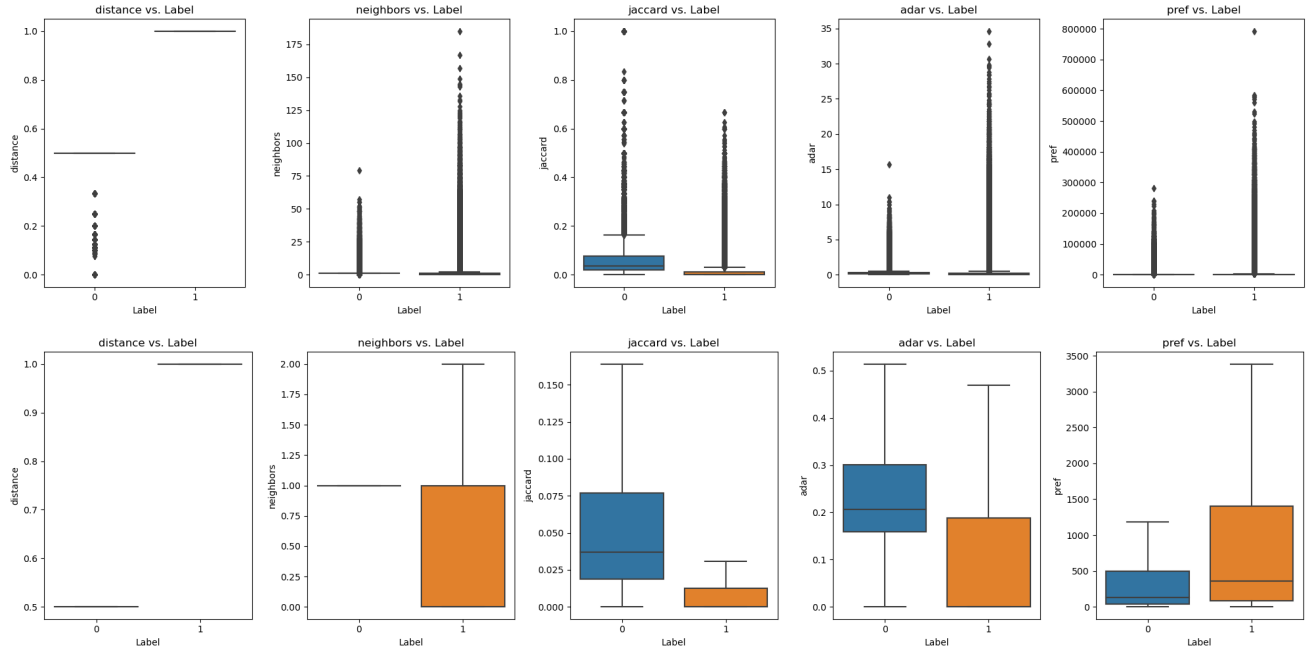
Σαν πρώτη προσέγγιση επιλέξαμε για κάθε πραγματική ακμή (u,v) να κρατάμε τον λιγότερο δημοφιλή κόμβο (source) και να τον συνδέουμε με κάποιον γείτονα του περισσότερο δημοφιλή κόμβου (target), ο οποίος δεν ανήκει στην γειτονιά του source. Όταν δεν υπάρχει κάποιος τέτοιος γείτονας τον συνδέουμε με κάποιον άλλον τυχαίο κόμβο από το V*. Γνωρίζοντας από την υπόθεση πως βρισκόμαστε σε ένα truncated scale free δίκτυο κατά το οποίο οι περισσότερο δημοφιλείς κόμβοι έχουν περισσότερες πιθανότητες να σχηματίσουν ακμή, προσπαθήσαμε με την προσέγγιση αυτή να δημιουργήσουμε ένα σύνολο δεδομένων κατά το οποίο:

- Οι μη υπαρκτές ακμές θα βρίσκονται κοντά στις πραγματικές ακμές κι έτσι ο ταξινομητής θα μάθει να συγκρίνει "κοντινές" ακμές σε όρους απόστασης. Αυτό είναι σημαντικό επειδή το V*xV* είναι ένας πολύ μεγάλος χώρος και αν επιλέγαμε όλες τις μη υπαρκτές ακμές με

τυχαίο τρόπο κατά βάση θα οδηγούμασταν σε πολύ μακρινές αποστάσεις μεταξύ των κόμβων.

- Οι μη υπαρκτές ακμές θα αποτελούνται κατά βάση από έναν λιγότερο δημοφιλή κόμβο και κάποιον άλλον που δεν γνωρίζουμε την δημοφιλία του, κι έτσι αναμένουμε να έχουμε μικρότερες τιμές σε Preferential Attachment από'τι στις υπαρκτές ακμές. Με αυτόν τον τρόπο θα οδηγήσουμε τον ταξινομητή να δίνει περισσότερο έμφαση στο Preferential Attachment για να ξεχωρίζει τις 2 κλάσεις και έτσι θα τον βοηθήσουμε να συμπεριλάβει την 2η ιδιότητα του scale-free δικτύου (Preferential Attachment).

Έπειτα από την δημιουργία του training & test dataset με την στρατηγική αυτή οι κατανομές των μέτρων ομοιότητας για κάθε label ξεχωριστά είχαν την εξής μορφή στο training dataset:

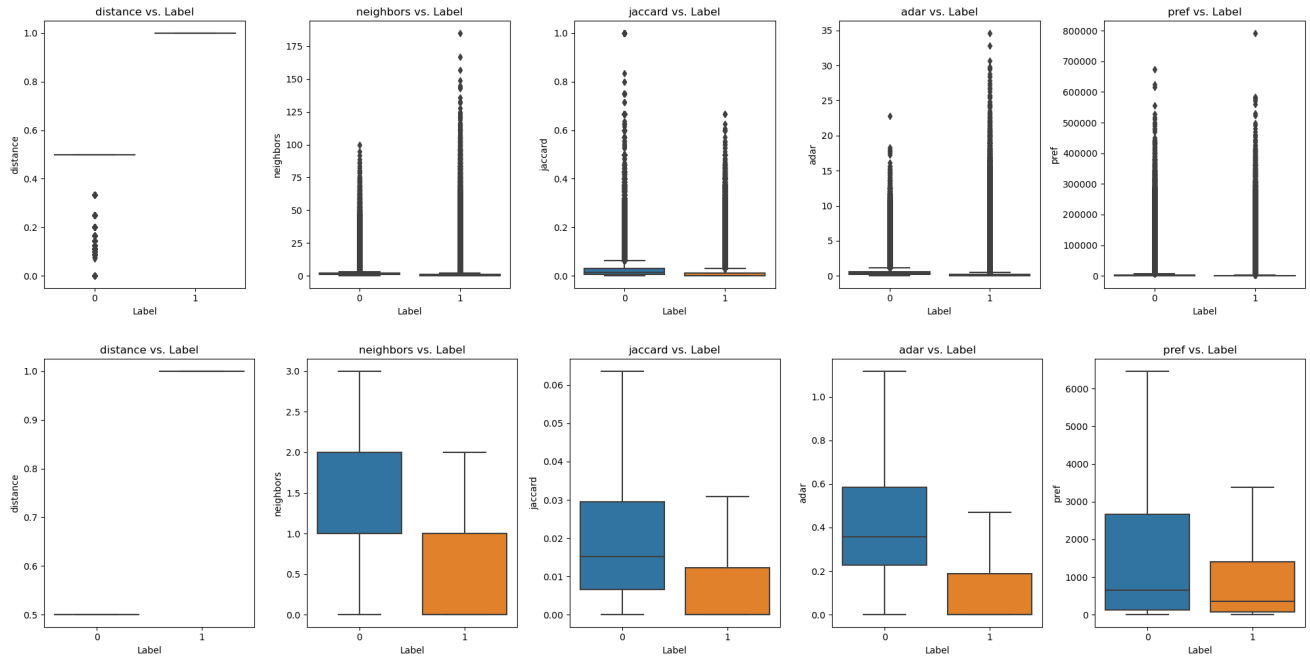


Παρατηρούμε πως όντως οι τιμές του preferential attachment των υπαρκτών ακμών τείνουν να είναι μεγαλύτερες ενώ στα υπόλοιπα μετρικά βλέπουμε πως οι μη υπαρκτές ακμές τείνουν να δίνουν λίγο μεγαλύτερες τιμές. Από την διαδικασία τις εκπαίδευσης και πρόβλεψης θα αφαιρέσουμε το μέτρο distance καθώς θα οδηγούμασταν σε τετριμμένες απαντήσεις από το μοντέλο μας, αφού όταν το Graph Distance = 1 , μας μαρτυράει πως οι 2 κόμβοι αυτοί είναι άμεσα συνδεδεμένοι μεταξύ τους άρα έχουν και υπαρκτή ακμή.

Έπειτα από την απαραίτητη προεπεξεργασία των δεδομένων προχωρήσαμε στην εκπαίδευση ενός νευρωνικού δικτύου με 3 επίπεδα και μιας απλής λογιστικής παλινδρόμησης. Το **test accuracy του νευρωνικού δικτύου έφτασε το 86.56%**, ενώ **το test accuracy της λογιστικής παλινδρόμησης έφτασε το 79.88%.**

2. Περισσότερο Δημοφιλής Κόμβος - Γείτονας Λιγότερο Δημοφιλή Κόμβου

Η δεύτερη στρατηγική που ακολουθήσαμε ήταν η αντίστροφη της 1ης. Αυτήν την φορά επιλέξαμε να κρατάμε σαν source τον δημοφιλέστερο κόμβο και να τον συνδέουμε με κάποιον γείτονα του λιγότερο δημοφιλή κόμβου. Η στρατηγική αυτή είχε σκοπό να εκπαιδεύσουμε έναν ταξινομητή ο οποίος θα μάθει να ξεχωρίζει υπαρκτές και μη υπαρκτές ακμές με βάση τα υπόλοιπα χαρακτηριστικά και όχι το preferential attachment. Έπειτα από την δημιουργία του training & test dataset με την στρατηγική αυτή οι κατανομές των μέτρων ομοιότητας για κάθε label ξεχωριστά είχαν την εξής μορφή στο training dataset:



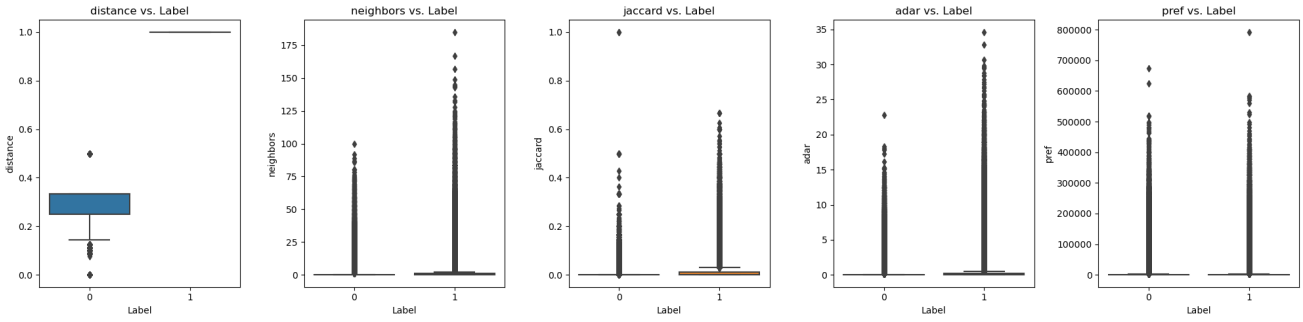
Παρατηρούμε πως με την προσέγγιση αυτή όλα τα μετρικά τείνουν να έχουν μεγαλύτερες τιμές στις μη υπαρκτές ακμές. Έπειτα από την απαραίτητη προεπεξεργασία των δεδομένων προχωρήσαμε στην εκπαίδευση ενός νευρωνικού δικτύου με 3 επίπεδα και μιας απλής λογιστικής παλινδρόμησης, όπως και στην προηγούμενη στρατηγική. Το **test accuracy του νευρωνικού δικτύου έφτασε το 81.28%**,

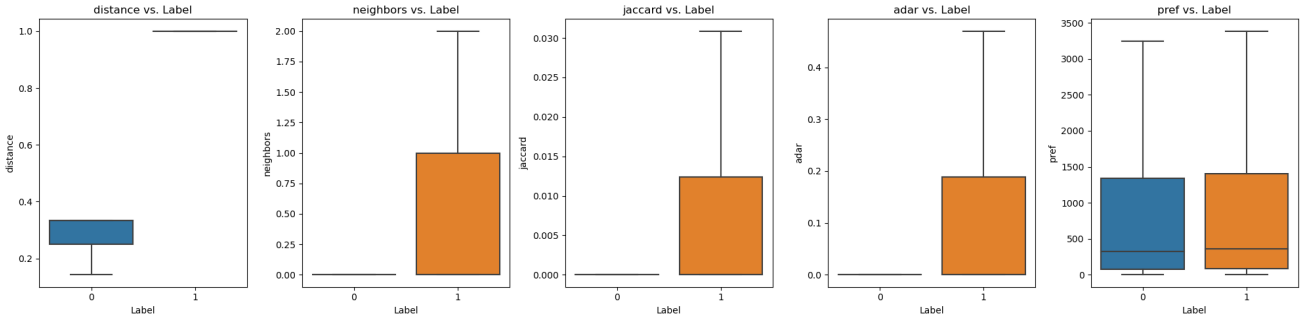
ενώ **το test accuracy της λογιστικής παλινδρόμησης έφτασε το 76.30%.** Βλέπουμε δηλαδή πως παίρνουμε χειρότερα αποτελέσματα.

3. Επιλογή μη υπαρκτών ακμών με βάση την δημοφιλία των κόμβων

Η τρίτη και τελευταία στρατηγική που ακολουθήσαμε ήταν να επιλέγουμε τα ζευγάρια κόμβων που θα εκπροσωπήσουν τις μη πιθανές ακμές με βάση την δημοφιλία τους. Επειδή βρισκόμαστε σε ένα δίκτυο όπου οι πιο δημοφιλής κόμβοι έχουν μεγαλύτερη πιθανότητα να αναπτύξουν κάποια ακμή από'τι οι λιγότερο δημοφιλείς κόμβοι, όρισαμε την πιθανότητα του να επιλεγέι κάποιος κόμβος για να σχηματίσει πιθανή ακμή ως p = degree / sum of degrees. Δηλαδή, όσο μεγαλύτερο είναι το degree ενός κόμβου τόσο μεγαλύτερη είναι η πιθανότητα να σχηματίσει ακμή. Με αυτόν τον τρόπο προσπαθήσαμε να προσομοιώσουμε τον τρόπο με τον οποίο αντλούμε τις μη υπαρκτές ακμές με τον τρόπο με τον οποίον δημιουργούνται οι υπαρκτές, κι έτσι να δημιουγήσουμε ένα σύνολο δεδομένων το οποίο θα έχει απαλλαχθεί πλήρως από την ιδιότητα του scale-free δικτύου (preferential attachment) . Για τον λόγο ότι το να υπολογίσουμε την πιθανότητα επιλογής ενός υποψήφιου απαιτείται μεγάλη υπολογιστική ισχύς, επιλέξαμε να διαμοιράσουμε τον κάθε γράφο σε batches ακμών και κόμβων δημιουργώντας κάθε φορά έναν Gbatch γράφο. Σε κάθε Gbatch υπολογίζαμε τις πιθανότητες επιλογής κάθε κόμβου με βάση την δημοφιλία του στον γράφο Gbatch και με βάση τις πιθανότητες αυτές, κάναμε τυχαία επιλογή ενός κόμβου για να δημιουγήσουμε μια μη υπαρκτή ακμή. Αναλυτικά η διαδικασία αυτή φαίνεται στον κώδικα που έχει επισυναπτεί με την παράδοση της εργασίας στο αρχείο *Version_3_Creation.ipynb*.

Δημιουργώντας λοιπόν το dataset πήραμε τις εξής κατανομές:





Βλέπουμε πως στην προσέγγιση αυτή έχουμε καταφέρει να ισορροπήσουμε τις κατανομές του preferential attachment κάτι που σηματοδοτεί πως έχουμε απαλλάξει τα δεδομένα μας από την ιδιότητα που προαναφέραμε. Όμως μπορούμε να διακρίνομε πως στα υπόλοιπα μέτρα ομοιότητας έχουμε κυρίως μηδενικές τιμές για της μη υπαρκτές ακμές. Έπειτα από την απαραίτητη προεπεξεργασία των δεδομένων προχωρήσαμε στην εκπαίδευση ενός νευρωνικού δικτύου με 3 επίπεδα και μιας απλής λογιστικής παλινδρόμησης, όπως και στις 2 προηγούμενες στρατηγικές. Το **test accuracy του νευρωνικού δικτύου έφτασε το 62.75%**, ενώ **το test accuracy της λογιστικής παλινδρόμησης έφτασε το 62.58%.** Βλέπουμε δηλαδή πως παίρνουμε τα χειρότερα αποτελέσματα από όλες τις στρατηγικές και αυτήν την φορά η λογιστική παλινδρόμηση και το νευρωνικό δίκτυο δίνουν πολύ κοντινά αποτελέσματα.

Συμπέρασμα

Παρατηρούμε πως την μεγαλύτερη ακρίβεια την έφερε η πρώτη προσέγγιση κατά την οποία καταφέραμε να αναδείξουμε στους ταξινομητές την διαφορά στις κατανομές του preferential attachment μεταξύ υπαρκτών και μη υπαρκτών ακμών. Παράλληλα, η χειρότερη επίδοση ήρθε από την τρίτη στρατηγική που καταφέραμε να ισορροπήσουμε τις κατανομές του preferential attachment μεταξύ υπαρκτών και μη υπαρκτών ακμών. Άρα, καταλαβαίνουμε πως επειδή βρισκόμαστε σε ένα δίκτυο όπου οι πιο δημοφιλείς κόμβοι έχουν μεγαλύτερη πιθανότητα να επιλεγούν, είναι υψίστης σημασίας να δημιουγήσουμε μη υπαρκτές ακμές με τέτοιον τρόπο, ώστε να αναδεικνύεται η διαφορά στις κατανομές του preferential attachment μεταξύ υπαρκτών και μη υπαρκτών ακμών κι έτσι ο ταξινομητής να συμπεριλάβει την ιδιότητα αυτή κατά την εκπαίδευσή του.