Χαροκόπειο Πανεπιστήμιο
Τμήμα Πληροφορικής και Τηλεματικής

Συστήματα Διαχείρισης Δεδομένων Μεγάλης Κλίμακας
Άσκηση

Διδάσκων	**Δημήτρης Μιχαήλ**
Ακ. Έτος	**2018-2019**
Ημ. Παράδοσης	**06/06/2019**

Οδηγίες Παράδοσης

1. Η άσκηση μπορεί να γίνει σε ομάδες των δύο φοιτητών ή ατομικά.
2. Η παράδοση της άσκησης πρέπει να γίνει ηλεκτρονικά μέσω της πλατφόρμας `http://eclass.hua.gr`. Μπορείτε να ανεβάσετε την άσκηση σας μέχρι και την ημέρα της παράδοσης.
3. Το παραπάνω zip αρχείο πρέπει να περιέχει
   (a) ένα φάκελο **src** με τον πηγαίο κώδικα της άσκησης
   (b) ένα .pdf αρχείο με την αναφορά.

   Το αρχείο πρέπει να περιέχει μόνο τον πηγαίο κώδικα και όχι και τα εκτελέσιμα αρχεία.
4. Η αναφορά πρέπει να περιέχει εισαγωγή στο θέμα, λεπτομερή ανάλυση της λύσης που υλοποιήσατε μαζί με τον κώδικα που γράψατε καθώς και παραδείγματα εκτέλεσης του.
5. Σε περίπτωση αντιγραφής θα μηδενίζονται όλες οι εμπλεκόμενες ασκήσεις.

Άσκηση

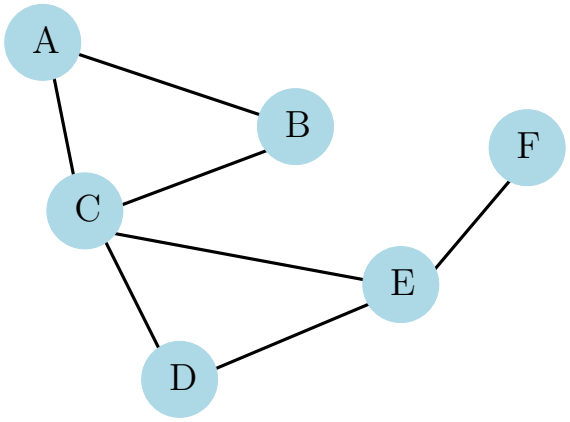
Στην άσκηση αυτή καλείστε να υπολογίσετε με το σύστημα Spark την μετρική *clustering coefficient* (συντελεστής ομαδοποίησης) για τους κόμβους ενός δικτύου. Η μετρική αυτή χρησιμοποιείτε ευρέως στην επιστήμη των δικτύων (Network Science) και μελετήθηκε αρχικά από τους Watts and Strogatz στο παρακάτω άρθρο τους στο περιοδικό Nature του 1998.

- Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. Nature, 393:440–442, 1998.

Το άρθρο μπορείτε να το βρείτε στον παρακάτω σύνδεσμο: `https://www.nature.com/articles/30918.pdf`

Clustering Coefficient

Σε ένα μη-κατευθυνόμενο γράφημα $G(V, E)$ όπου V είναι το σύνολο κόμβων και E το σύνολο των ακμών ο *συντελεστής ομαδοποίησης* ενός κόμβου v περιγράφει το ποσοστό των γειτόνων του v που είναι και αυτοί γείτονες μεταξύ τους. Στην περίπτωση των κοινωνικών δικτύων φιλίας για παράδειγμα, ο συντελεστής ομαδοποίησης ενός χρήστη είναι το ποσοστό των φίλων του που είναι και αυτοί μεταξύ τους φίλοι.



Ας πάρουμε για παράδειγμα τον κόμβο C στο παραπάνω σχήμα. Ο κόμβος C έχει βαθμό $d_C = 4$. Θυμηθείτε πως ο βαθμός (degree) ενός κόμβου σε μη-κατευθυνόμενα γραφήματα είναι ο αριθμός των γειτόνων. Ο κόμβος C αφού έχει d_C γείτονες, έχει $\binom{d_C}{2} = \frac{d_C(d_C-1)}{2} = 6$ ζεύγη γειτόνων. Από τα 6 αυτά ζεύγη μόνο τα 2 ζεύγη ((A, B) και (D, E)) είναι συνδεδεμένα με ακμή. Ο συντελεστής ομαδοποίησης για τον κόμβο C είναι ίσος με $\frac{2}{6} = \frac{1}{3}$.

Αντίστοιχα ο κόμβος E έχει συντελεστή ομαδοποίησης $\frac{1}{3}$ ενώ οι A, B και D έχουν συντελεστή ίσο με 1 αφού έχουν μόνο ένα ζεύγος γειτόνων και αυτό το ζεύγος είναι συνδεδεμένο με ακμή.

Ιδιαίτερη περίπτωση αποτελούν όσοι κόμβοι έχουν 1 ή 0 γείτονες, όπως για παράδειγμα ο κόμβος F. Σε αυτή την περίπτωση ορίζουμε τον συντελεστή ίσο με μηδέν.

Global Clustering Coefficient

Για τον υπολογισμό μίας καθολικής τιμής για το γράφημα υπάρχουν δύο ορισμοί: είτε (α) χρησιμοποιώντας τον μέσο όρο των τιμών των κόμβων, είτε (β) χρησιμοποιώντας τριάδες. Στην παρούσα άσκηση θα χρησιμοποιήσουμε την πρώτη μέθοδο υπολογισμού όπου ο καθολικός συντελεστής ομαδοποίησης ορίζεται ως ο μέσος όρος των συντελεστών ομαδοποίησης των κόμβων.

Στο παράδειγμα του σχήματος είναι

$$\frac{1+1+\frac{1}{3}+1+\frac{1}{3}+0}{6} = \frac{11}{18}$$

Spark

Χρησιμοποιήστε το Spark για να υπολογίσετε τον συντελεστή ομαδαποίησης για κάθε κόμβο ενός γραφήματος καθώς και τον καθολικό συντελεστή. Θα πρέπει να χρησιμοποιήσετε ένα γράφημα από την συλλογή SNAP (`https://snap.stanford.edu/data/index.html`). Φροντίστε το γράφημα που θα διαλέξετε να είναι μη-κατευθυνόμενο (undirected) και μέτριου μεγέθους ώστε να μπορεί να εκτελεστεί στον υπολογιστή σας.

Προσοχή για να γίνει δεκτή η άσκηση σας θα πρέπει να χρησιμοποιεί το Spark σε επίπεδο RDD και όχι να χρησιμοποιεί κάποια πιο υψηλού επιπέδου υλοποίηση όπως π.χ τα Mllib ή GraphX. Με άλλα λόγια, δεν επιτρέπεται να χρησιμοποιήσετε βιβλιοθήκη του Spark πλέον της κύριας (core).

Το πρόγραμμα σας πρέπει να δέχεται από την γραμμή εντολών ως είσοδο το όνομα του αρχείου που περιέχει το γράφημα. Στην έξοδο θα πρέπει να δίνει ζεύγη της μορφής

```
(graph-vertex, clustering-coefficient)
```

καθώς και να τυπώνει στην τυπική έξοδο την καθολική τιμή του clustering coefficient.

Βαθμολογία

Βαθμολογήστε για:

- Καλή μοντελοποίηση, χρήση λίγων RDD και σωστό caching των ενδιάμεσων αποτελεσμάτων.

- Σωστή ονοματολογία μεταβλητών και συναρτήσεων.
- Σωστή λειτουργικότητα.
- Αποδοτική υλοποίηση.
- Ολοκληρωμένη και σωστή τεκμηρίωση και περιγραφή στην αναφορά.

Μπορειτε να χρησιμοποιήστε (a) Java και Maven όπως στις εργαστηριακές ασκήσεις ή (b) Python. Μπορείτε να χρησιμοποιήσετε ως σκελετό τον κώδικα του εργαστηρίου. Εκτός από τα αρχεία με τον κώδικα πρέπει να γράψετε και μια αναφορά μερικών σελίδων. Η αναφορά πρέπει να εξηγεί τις διάφορες επιλογές που κάνατε, γιατί μοντελοποιήσατε έτσι το πρόβλημα καθώς και να σχολιάζει τον κώδικα σας. Η αναφορά πρέπει να είναι υποχρεωτικά σε μορφή *pdf*.