Θέματα Διπλωματικών Εργασιών
Εαρινό Εξάμηνο Ακαδημαϊκού Έτους 2020-2021
(και Χειμερινού Εξαμήνου Ακαδημαϊκού Έτους 2021-2022)

- Ακολουθεί μια λεπτομερή περιγραφή των θεμάτων των διπλωματικών εργασιών για τα δύο επόμενα εξάμηνα. Τα θέματα θα ανανεωθούν το Μάϊο του 2021.
- Διαβάστε τις σχετικές περιγραφές και αν σας ενδιαφέρει κάποιο θέμα στείλτε μου email ***μέχρι και τις 15.1.2021*** για όσους ενδιαφέρονται για το ***εαρινό εξάμηνο***. Όσοι ενδιαφέρονται για του χρόνου μπορεί να επικοινωνήσουν μαζί μου και αργότερα.
- Στο email να αναφέρετε ποιο θέμα σας ενδιαφέρει και επισυνάψτε αν είναι δυνατόν αναλυτική βαθμολογία.

Διπλωματική 1: Δίκαιος τυχαίος περίπατος (fair random walk)

Ο τυχαίος περίπατος είναι ένας κλασικός τρόπος διάσχισης ενός γράφου. Με απλά λόγια, ξεκινάμε από ένα τυχαίο κόμβο, επισκεπτόμαστε τυχαία κάποιο γείτονα του και μετά τυχαία κάποιο γείτονά του γείτονα κοκ. Οι τυχαίοι περίπατοι αποτελούν βασικό κομμάτι πολλών αλγορίθμων όπως ο PageRank και μιας σημαντικής κατηγορίας graph embedding τεχνικών (τεχνικών μηχανικής μάθησης για αναπαράσταση κόμβων ως διανύσματα).

Σε αυτή τη διπλωματική θα μελετηθούν δίκαιες εκδοχές των τυχαίων περιπάτων. Συγκεκριμένα, σε πολλά δίκτυα (ιδιαίτερα κοινωνικά δίκτυα, όπου οι κόμβοι του δικτύου είναι πρόσωπα και οι συνδέσεις απεικονίζουν σχέσεις, όπως φιλία, συνεργασία, επικοινωνία κλπ), οι κόμβοι χαρακτηρίζονται από ευαίσθητα χαρακτηριστικά, όπως η ηλικία, το γένος, ή το φύλλο τους. Σε τέτοιες περιπτώσεις θα θέλαμε ο τυχαίος περίπατος να είναι δίκαιος, δηλαδή, διαισθητικά, θα θέλαμε να επισκέπτεται κόμβους με όλα τα χαρακτηριστικά (π.χ., και άντρες και γυναίκες).

Σε προηγούμενη δουλειά μας (στην εργασία [1]) έχουμε προτείνει μια δίκαιη εκδοχή του τυχαίου περιπάτου σε συνδυασμό με το PageRank [2].

Ο αλγόριθμος PageRank είναι ίσως ο πιο γνωστός αλγόριθμος υπολογισμού της σημαντικότητας ενός κόμβου στο δίκτυο και ήταν ο βασικός αλγόριθμος πίσω από την επιτυχία της Google. ***Καθώς τα κοινωνικά δίκτυα συνεχώς μεγαλώνουν και η επιρροή των χρηστών (κόμβων) που συμμετέχουν σε αυτά αυξάνεται, είναι ανάγκη να έχουμε αξιόπιστους και δίκαιους τρόπους να αξιολογούμε τη σημαντικότητά τους,***

Στην διπλωματική αυτή θα μελετηθούν παραλλαγές της εργασίας [1]. Στην [1], όλοι οι κόμβοι ενός δικτύου είναι δίκαιοι, δηλαδή, κάνουν δίκαιους περιπάτους. Στη διπλωματική:

(α) θα εξετάσουμε την περίπτωση όπου μόνο συγκεκριμένα υποσύνολα κόμβων κάνουν δίκαιους τυχαίους περιπάτους, ή/και (β) τη αντικατάσταση του τυχαίου περίπατου σε γνωστούς αλγορίθμους embedding με το δίκαιο τυχαίο περίπατο.

Θα χρειαστεί:
(1) Θεωρητική κατανόηση του PageRank
(2) Υλοποίηση της παραλλαγής – υπάρχει ήδη υλοποίηση του βασικού δίκαιου περίπατου (σε C και Python), και
(3) Πειραματική αξιολόγηση με χρήση συνθετικών και πραγματικών δικτύων (αυτά τα δίκτυα τα έχουμε ήδη δημιουργήσει/συλλέξει). Τα δίκτυα που θα χρησιμοποιηθούν αφορούν πραγματικά κοινωνικά δίκτυα και οι μετρήσεις θα δείξουν κατά πόσο είναι δίκαια στην παρούσα μορφή τους και κατά πόσο ο νέος δίκαιος αλγόριθμος τα βελτιώνει.

Αναφορές
[1] Sotiris Tsioutsiouliklis, Evaggelia Pitoura, Panayiotis Tsaparas, Ilias Kleftakis, Nikos Mamoulis: Fairness-Aware Link Analysis. CoRR abs/2005.14431 (2020)
https://arxiv.org/abs/2005.14431
(υπάρχει ενημερωμένη εργασία, αλλά και το [1] έχει τα βασικά, μας ενδιαφέρουν οι local algorithms)
[2] Περιγραφή του PageRank υπάρχει σε πολλά βιβλία, αλλά και στο Wikipedia. Δείτε το Κεφάλαιο 5 του βιβλίου: Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman: Mining of Massive Datasets, 2nd Ed. Cambridge University Press 2014
http://infolab.stanford.edu/~ullman/mmds/ch5.pdf

Διπλωματική 2: Δίκαια συστήματα συστάσεων με χρήση του αλγορίθμου SALSA

Τα συστήματα συστάσεων προτείνουν στους χρήστες αντικείμενα (όπως, ταινίες, εστιατόρια, βιβλία, δουλειές, κλπ) που πιθανών να τους ενδιαφέρουν. Η χρήση τους είναι ευρεία, όπως συστάσεις πιθανών συνδέσεων στα κοινωνικά δίκτυα, τηλεοπτικών σειρών στα συνδρομητικά κανάλια, θέσεων επαγγελματικής απασχόλησης, εργαζομένων, κοκ

Τα συστήματα συστάσεων συνήθως χρησιμοποιούν ιστορική πληροφορία, όπως ποια αντικείμενα ο ίδιος ο χρήστης ή χρήστες παρόμοιοι με αυτόν έχουν διαλέξει (δηλαδή, αγοράσει, δει, αξιολογήσει, κλπ) στο παρελθόν [1]. Σε αυτήν την εργασία θα εξετάσουμε έναν αλγόριθμο συστάσεων που ονομάζεται SALSA [2, 3] και βασίζεται σε ένα διμερή γράφο ανάμεσα στους χρήστες και στα αντικείμενα. Με απλά λόγια, σε αυτό το γράφο, οι κόμβοι αντιστοιχούν στους χρήστες και στα αντικείμενα και υπάρχει μια ακμή από ένα χρήστη σε ένα αντικείμενο, αν ο χρήσης έχει διαλέξει το συγκεκριμένο αντικείμενο στο παρελθόν. Ο SALSA κάνει κάποιους τυχαίους περιπάτους σε αυτό το γράφο. Ο τυχαίος περίπατος είναι ένας κλασικός τρόπος διάσχισης ενός γράφου. Με απλά λόγια, ξεκινάμε από ένα τυχαίο κόμβο, επισκεπτόμαστε τυχαία κάποιο γείτονα του και μετά τυχαία κάποιο γείτονά του γείτονα κοκ.

Σε αυτή τη διπλωματική θα μελετηθούν δίκαιες εκδοχές του αλγορίθμου SALSA. ***Θα θέλαμε οι συστάσεις να είναι δίκαιες τόσο για τους χρήστες όσο και για τα αντικείμενα. Για παράδειγμα, θα θέλαμε μια τεχνολογική δουλειά να προτείνεται και σε άντρες και σε γυναίκες. Αυτό είναι ένα πρόβλημα με μεγάλη σημασία μιας και η εξάρτησή μας στις συστάσεις αυξάνεται συνεχώς και καθορίζει πολλές από τις αποφάσεις μας***.

Σε προηγούμενη δουλειά μας (στην εργασία [4]) έχουμε προτείνει μια δίκαιη εκδοχή του τυχαίου περιπάτου. Στη διπλωματική, θα μελετηθεί πως η βασική ιδέα αυτού του αλγορίθμου μπορεί να χρησιμοποιηθεί στο SALSA ώστε να έχουμε δίκαιες συστάσεις.

Θα χρειαστεί:
(1) Θεωρητική κατανόηση των αλγορίθμων συστάσεων και του SALSA
(2) Υλοποίηση του SALSA και σχεδιασμός και υλοποίηση της παραλλαγής του
(3) Πειραματική αξιολόγηση με χρήση πραγματικών δεδομένων συστάσεων (όπως το movieLens dataset και άλλα).

Αναφορές
[1] Περιγραφή για τα συστήματα συστάσεων υπάρχει σε πολλά βιβλία. Για να πάρετε μια ιδέα δείτε το Κεφάλαιο 9 του βιβλίου: Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman: Mining of Massive Datasets, 2nd Ed. Cambridge University Press 2014
http://infolab.stanford.edu/~ullman/mmds/ch9.pdf

[2] Ronny Lempel, Shlomo Moran: SALSA: the stochastic approach for link-structure analysis. 131-160 (η εργασία που πρότεινε τον αλγόριθμο, η εργασία είναι κάπως δυσνόητη, η βασική ιδέα είναι απλή)
http://denif.enslyon.fr/data/algorithmique_reseaux_telecoms_enslyon/2006/tds/SALSA.pdf

[3] Pankaj Gupta, Ashish Goel, Jimmy J. Lin, Aneesh Sharma, Dong Wang, Reza Zadeh: WTF: the who to follow service at Twitter. WWW 2013: 505-514 (στη εργασία αυτή χρησιμοποιούν το SALSA για συστάσεις στο Twitter, στο Section 5.2 είναι μια σύντομη και πιο εύκολα κατανοητή περιγραφή του SALSA)
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.699.1726&rep=rep1&type=pdf

[4] Sotiris Tsioutsiouliklis, Evaggelia Pitoura, Panayiotis Tsaparas, Ilias Kleftakis, Nikos Mamoulis: Fairness-Aware Link Analysis. CoRR abs/2005.14431 (2020)
https://arxiv.org/abs/2005.14431
(υπάρχει ενημερωμένη εργασία, αλλά και το [1] έχει τα βασικά, μας ενδιαφέρουν οι local algorithms)

Διπλωματική 3: Χρήση embeddings για συστάσεις σε συστήματα γραφοβάσεων (graph databases)

Τα συστήματα γραφοβάσεων διαφέρουν από τα σχεσιακά συστήματα διαχείρισης βάσεων δεδομένων στο ότι το μοντέλο δεδομένων που χρησιμοποιούν δεν είναι το σχεσιακό μοντέλο, αλλά το μοντέλο ενός γράφου με ιδιότητες (property graphs). Με απλά λόγια τα δομικά στοιχεία του μοντέλου είναι (α) κόμβοι που έχουν ιδιότητες και (β) ακμές που εκφράζουν τις σχέσεις μεταξύ των κόμβων. Η γλώσσα ερωτήσεων διαφέρει από την SQL στο ότι επιτρέπει διασχίσεις μονοπατιών και εύρεση συγκεκριμένων προτύπων. Στα πλαίσια αυτής της διπλωματικής, θα χρησιμοποιηθεί η πιο γνωστή γραφοβάση η Neo4j [1].

Στόχος είναι η επέκταση των γραφοβάσεων με συστάσεις. Συγκεκριμένα, ο χρήστης θα κάνει ένα ερώτημα στη βάση δεδομένων. Με βάση το ερώτημα και την απάντηση σε αυτό το ερώτημα, το σύστημα θα του προτείνει επιπρόσθετα αποτελέσματα τα οποία δεν

ανήκουν στο αποτέλεσμα του αρχικού του ερωτήματος αλλά μπορεί επίσης να τον ενδιαφέρουν [2].

Για τη δημιουργία των συστάσεων θα εξετάσουμε graph embeddings [3]. Τα graph embeddings χρησιμοποιούν μηχανική μάθηση για να απεικονίσουν τους κόμβους ενός δικτύου σε πολύ-διάστατα διανύσματα. Η βασική ιδέα αυτών των απεικονίσεων είναι παρόμοιοι κόμβοι (πχ, κόμβοι με πολλούς κοινούς γείτονες) να απεικονίζονται σε παρόμοια διανύσματα.

Αυτό είναι ένα θεμελιακό πρόβλημα καθώς με την μεγάλη ανάπτυξη της μηχανικής μάθησης, η ενσωμάτωση της σε συστήματα διαχείρισης δεδομένων είναι ένα βασικό ζητούμενο.

Η βασική ιδέα της διπλωματικής είναι να χρησιμοποιήσουμε τα embeddings για να αναπαραστήσουμε τα αποτελέσματα της ερώτησης. Με βάση αυτήν την αναπαράσταση θα προτείνουμε στο χρήστη δεδομένα των οποίων τα embeddings μοιάζουν με αυτά του αποτελέσματος.

Θα χρειαστεί:
(1) Κατανόηση των γραφοβάσεων και των graph embedding
(2) Υλοποίηση στη Neo4j (για τους αλγορίθμους για graph embeddings θα χρησιμοποιήσουμε τις διαθέσιμες υλοποιήσεις τους)
(3) Πειραματική αξιολόγηση

Αναφορές
[1] https://neo4j.com/
[2] Kostas Stefanidis, Marina Drosou, Evaggelia Pitoura, You May Also Like Results in Relational Databases, In 3rd International Workshop on Personalized Access, Profile Management, and Context Awareness in Databases (PersDB 2009)
(περιγράφει κάποιες ιδέες για σχεσιακές βάσεις δεδομένων)
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.297.4011&rep=rep1&type=pdf
[3] Aditya Grover, Jure Leskovec: node2vec: Scalable Feature Learning for Networks. KDD 2016: 855-864
(ένας από τους πολλούς αλγορίθμους για node embeddings)
https://snap.stanford.edu/node2vec/

Διπλωματική 4: Εύρεση συνεκτικών πυκνών γραφημάτων σε χρονικά μεταβαλλόμενα γραφήματα

Πολλά από τα πραγματικά δίκτυα (κοινωνικά, υπολογιστικά, βιολογικά, κλπ) μεταβάλλονται στο χρόνο. Σε αυτήν την διπλωματική εργασία θέλουμε να βρούμε εκείνα τα σύνολα κόμβων που μένουν ισχυρά συνδεδεμένα μεταξύ τους στο χρόνο. Αυτά τα σύνολα μπορεί να αντιστοιχούν για παράδειγμα σε ομάδες χρηστών που επικοινωνούν μεταξύ σε ένα κοινωνικό δίκτυο ή συνεργάζονται μεταξύ τους σε όλες (ή σχεδόν όλες) τις χρονικές περιόδους, δηλαδή θέλουμε τις σχέσεις που διατηρούνται στο χρόνο.

Σε προηγούμενη δουλειά μας [1] επικεντρωθήκαμε στον εντοπισμό των πυκνά συνδεδεμένων κόμβων, δηλαδή, των κόμβων που μεταξύ τους υπάρχουν πολλές συνδέσεις σε όλες (ή σχεδόν όλες) τις χρονικές περιόδους. Ο προτεινόμενος αλγόριθμος είναι σχετικά απλός και βασίζεται στην επαναληπτική αφαίρεση του κόμβου με το μικρότερο βαθμό.

Ο σκοπός αυτής της εργασίας είναι διττός: (1) να προσθέσει την απαίτηση αυτοί οι κόμβοι να είναι συνδεδεμένοι και (2) να εφαρμόσει τον αλγόριθμο σε νέα σύνολα πραγματικών δεδομένων. ***Γενικός στόχος είναι να κάνει τον προτεινόμενο αλγόριθμο ένα πρακτικό εργαλείο για την μελέτη χρονικά μεταβαλλόμενων γραφημάτων. Αυτό είναι πολύ σημαντικό γιατί η μελέτη των χρονικά μεταβαλλόμενων γραφημάτων είναι στο κέντρο του επιστημονικού ενδιαφέροντος και δεν υπάρχουν πολλά σχετικά εργαλεία.***

Θα χρειαστεί:
(1) Κατανόηση του βασικού αλγορίθμου
(2) Τροποποίηση του (υπάρχει υλοποίηση σε Java)
(3) Πειραματική αξιολόγηση σε υπάρχοντα χρονικά γραφήματα και σε γραφήματα που θα συλλεχθούν στα πλαίσια της διπλωματικής όπως δημιουργία ενός δικτύου με δεδομένα από το twitter, όπου κόμβοι είναι τα hashtags και υπάρχει ακμή μεταξύ δύο κόμβων αν αυτοί εμφανίζονται στο ίδιο tweet).

Αναφορές
[1] Konstantinos Semertzidis, Evaggelia Pitoura, Evimaria Terzi, Panayiotis Tsaparas: Finding lasting dense subgraphs. Data Min. Knowl. Discov. 33(5): 1417-1445 (2019) http://cs.uoi.gr/~ksemer/docs/publications/pkdd19.pdf