Πολυτεχνική Σχολή
Τμήμα Μηχανικών Η/Υ και Πληροφορικής

Μέθοδοι Μητρώων και Υπολογιστικά Εργαλεία στην Επιστήμη Δεδομένων
ΥΔΑ

Ranking Hubs and Authorities Using Matrix Functions

Κωνσταντίνος Πετράκης
Α.Μ. 1041589

Μάρτιος 2021

1 Εισαγωγή

Τα τελευταία χρόνια η έρευνα γύρω απο το πεδίο των δικτύων έχει επεκταθεί σημαντικά έξω απο τον χώρο των μαθηματικών σε πολλούς επιστημονικούς τομείς εξαιτίας της ευρείας εφαρμογής που έχουν τα δίκτυα για την μοντελοποίηση διάφορων προβλημάτων, από τα κοινωνικά δίκτυα και τα δίκτυα αλληλεπίδρασης πρωτεινών μέχρι την οικονομία. Συνήθως όμως σε αυτά τα δίκτυα οι κόμβοι είναι ετερογενής, παρουσιάζοντας πολύ διαφορετικούς ρόλους στη δομή και τη λειτουργία του δικτύου. Ο εντοπισμός λοιπόν των σημαντικών κόμβων είναι υψίστης σημασίας και είναι η επιτυχημένη ανίχνευση αυτών των κρίσιμων κόμβων που επιτρέπει, μεταξύ άλλων, τον έλεγχο στο ξέσπασμα επιδημιών, τον εντοπισμό των χρηστών με την μεγαλύτερη επιρροή σε ένα κοινωνικό δίκτυο και την ανακάλυψη βασικών πρωτεινών.

Οι μετρικές που χρησιμοποιούνται για την αξιολόγηση της σπουαιδότητας ενός κόμβου αναφέρονται συνήθως σαν μέτρα κεντρικότητας. Ένα μέτρο κεντρικότητας αναθέτει μία τιμή σε κάθε κόμβο του δικτύου με τέτοιο τρόπο ώστε οι τιμές που παράγονται να παρέχουν μια κατάταξη των κόμβων ανάλογα με την σημαντικότητα τους. Εξ' αιτίας των πολλών ερμηνείων που μπορεί να έχει η έννοια της σημαντικότητας έχουν προταθεί πολλές διαφορετικές τεχνικές.

Υπάρχουν δύο μεγάλες κατηγορίες μεθόδων για την αναγνώριση σημαντικών κόμβων. Αυτές που χρησιμοποιούν μόνο δομική πληροφορία του δικτύου (π.χ degree-centrality, Katz centrality, subgraph centrality) και οι μέθοδοι επαναληπτικής βελτιώσης για την κατάταξη των κόμβων με βάση την σημαντικότητα τους. Βέβαια και οι μέθοδοι επαναληπτικής βελτίωσης κάνουν χρήση της δομής του δικτύου. Απλά επειδή αυτές οι δομικές ιδιότητες εξερευνώνται μέσω δυναμικών διεργασιών, όπως οι τυχαίοι περίπατοι, οι τεχνικές αυτές κατατάσσονται σε αυτή τη ξεχωριστή κατηγορία. Σε αυτή την δεύτερη κατηγορία τεχνικών η επιρροή ενός κόμβου δεν καθορίζεται μόνο απο το πλήθος των γειτόνων του αλλά και απο την επιρροή αυτών. Παραδείγματα τέτοιων τεχνικών αποτελούν οι αλγόριθμοι PageRank και HITS.

Κύριο αντικείμενο της παρούσας εργασία αποτελούν η κεντρικότητα υπογράφου (υπολογισμένη χρησιμοποιώντας το εκθετικό ενός μητρώου), ο αλγόριθμος HITS και η σχέση που συνδέει τις δύο αυτές τεχνικές. Στη συνέχεια στην ενότητα 2 παρουσιάζεται η βασική ορολογία που θα χρησιμοποιηθεί, διατυπώνεται πιο αυστηρά το πρόβλημα που μελετάμε και σκιαγραφούνται οι τεχνικές προσέγγισης του προβλήματος που θα συγκριθούν. Στην ενότητα 3 παρουσιάζονται τα βασικά στοιχεία της υπό μελέτη εργασίας [1] της οποίας στόχος είναι η επέκταση κάποιων καλά ορισμένων μέτρων κεντρικότητας, ειδικότερα της κεντρικότητας υπογράφου (subgraph centrality), απο τα μη-κατευθυνόμενα στα κατευθυνόμενα δίκτυα, μέσω της μοντελοποίησης αυτών των κατευθυνόμενων δικτύων με ένα διμερές (bipartite) γράφο. Τα μέτρα κεντρικότητας που εισάγονται εδώ επεκτείνουν τον τρόπο με τον οποία κατατάσσονται οι κόμβοι ενός δικτύου σε hubs και authorities μέσω του αλγορίθμου HITS με τον ίδιο τρόπο που η κεντρικότητα υπογράφου και η κεντρικότητα Katz μπορούν να θεωρηθούν σαν επέκταση της κεντρικότητας που ορίζεται με βάση τα ιδιοδιανύσματα.

2 Βασικές Έννοιες και Πρόβλημα Μελέτης

Εδώ θα αναφερθούμε με λίγο περισσότερη λεπτομέρεια στις τεχνικές που αναφέρθηκαν στην ενότητα 1 και την εφαρμογή τους στα μη-κατευθυνόμενα δίκτυα ώστε να γίνει ευκολότερα κατανοητή η επέκταση τους στα κατυεθυνόμενα δίκτυα στην επόμενη ενότητα. Όπως θα γίνει εμφανές παρακάτω τα μέτρα κεντρικότητας ορίζονται με τη χρήση διάφορων συναρτήσεων μητρώων. Τα περισσότερα μέτρα κεντρικότητας εφαρμόζονται εξίσου σε κατευθυνόμενα και μη-κατευθυνόμενα δίκτυα. Στα κατευθυνόμενα δίκτυα ωστόσο κάθε κόμβος έχει δύο ρόλους, είτε πηγή είτε δέκτης, καθώς μπορεί να είναι η αρχή ή το τέλος ενός συνόλου κατευθυνόμενων ακμών. Αυτός ο ξεχωριστός ρόλος των κόμβων ενός κατευθυνόμενου δικτύου οδήγησε στις έννοιες hubs και authorities, όπως θα δούμε παρακάτω.

Αρχικά για τη μη-κετυθυνόμενη περίπτωση, συμβολίζοντας με A το μητρώο γειτνίασης του μη-κατευθυνόμενου δικτύου, η κεντρικότητα υπογράφου για κάθε κόμβο i του δικτύου λαμβάνει με κάποιο τρόπο υπ' όψιν το πλήθος των υπογράφων στους οποίους συμμετέχει ο κόμβος i και είναι ίση με $[e^A]_{ii}$. Η συνάρτηση εκθετικού του μητρώου A δίνεται από το ανάπτυγμα της σειράς

$$e^A = I + A + \frac{A^2}{2!} + \frac{A^3}{3!} + ... + \frac{A^k}{k!} + ...$$

και λαμβάνοντας υπ' όψιν το γεγονός πως το στοιχείο $[A^k]_{ij}$ μετράει το πλήθος των περιπάτων μήκους k μεταξύ των κόμβων i και j, μπορούμε να ορίσουμε την κεντρικότητα υπογράφου [2] για κάθε κόμβο i σαν το συνολικό αριθμό των κλειστών περιπάτων που εκκινούν και τελειώνουν στον κόμβο i, ζυγίζοντας το πλήθος των περιπάτων μήκους k με τον παράγοντα $\frac{1}{k!}$. Τα μη διαγώνια στοιχεία του εκθετικού του μητρώου, $[e^A]_{ij}$, αντιστοιχούν στο πλήθος των περιπάτων μεταξύ των κόμβων i και j κλιμακώνοντας και πάλι τους περιπάτους μήκους k με τον παράγοντα $\frac{1}{k!}$. Αυτή η ποσότητα ονομάζεται δυνατότητα επικοινωνίας (communicability) και αποτελεί μέτρο για το πόσο εύκολα ρέει πληροφορία μεταξύ δύο κόμβων. Με απλά λόγια η κεντρικότητα υπογράφου δείχνει πόσο καλά συνδεδεμένος είναι ένας κόμβος με το υπόλοιπο δίκτυο. Συπληρωματικά μια εναλλακτική ερμηνεία της κεντρικότητας υπογράφου δίνεται κανονικοποιώντας κάθε στοιχείο του εκθετικού του μητρώου γειτνίαση με το ίχνος αυτού. Κοιτώντας στα διαγώνια στοιχεία του μητρώου $e^A/Tr(e^A)$ τώρα, η κεντρικότητα υπογράφου είναι ανάλογη της πιθανότητας να βρούμε έναν τυχαίο περιπατητή (random walker) κοντά στον κόμβο i. Η πιθανότητα αυτή είναι ίση με την πιθανότητα να επιλεγεί οποιοσδήποτε κλειστός περίπατος που ξεκινά απο τον κόμβο i και καταλήγει στον κόμβο i, απο το σύνολο όλων των κλειστών περιπάτων που εκκινούν και καταλήγουν στον ίδιο κόμβο, για όλους τους κόμβους του δικτύου.

Με παρόμοιο τρόπο, αντικαθιστώντας την συνάρτηση εκθετικού με την συνάρτηση resolvent, η οποία ορίζεται ως

$$R(A;c) = (I - cA)^{-1} = I + cA + c^2A^2 + ... + c^kA^k + ...$$

όπου $1 < c < \frac{1}{\lambda_{max}(A)}$, ώστε η σειρά να συγκλίνει και λαμβάνοντας τα διαγώνια στοιχεία του μητρώου $[(I - cA)^{-1}]_{ii}$, προκύπτει ένας εναλλακτικός τρόπος υπολογισμού την κεντρικότητας υπογράφου. Η διαφορά με την συνάρτηση εκθετικού

έγκειται στο γεγονός πως με τη χρήση της συνάρτησης resolvent λαμβάνονται περισσότερο υπ' οψιν, έχουν μεγαλύτερο βάρος, περίπατοι μικρότερου μήκους, με τους περιπάτους μεγαλύτερου μήκους να έχουν ελάχιστη επιρροή στην κεντρικότητα ενός κόμβου. Επίσης το εκθετικό δεν απαιτεί τον καθορισμό καμίας παραμέτρου, σε αντίθεση με την συνάρτηση resolvent τα αποτελέσματα της οποία εξαρτώνται σε μεγάλο βαθμό απο την ρύθμιση της παραμέτρου c. Με βάση την συνάρτηση resolvent ορίζεται και η κεντρικότητα Katz για κάθε κόμβο i ως εξής

$$K_i(c) = [(I - cA)^{-1}\mathbf{1}]_i$$

όπου $\mathbf{1}$ το διάνυσμα όλο άσους. Δηλαδή η κεντρικότητα Katz ορίζεται σαν το άθροισμα των γραμμών του μητρώου γειτνίασης A.

Ίσως ο πιο γνωστός αλγόριθμος για την εύρεση σημαντικών κόμβων σε κατευθυνόμενα δίκτυα είναι ο PageRank, σύμφωνα με τον οποίο η σπουδαιότητα ενός κόβμου καθορίζεται εξίσου απο την ποσότητα αλλά και απο την ποιότητα των κόμβων που δείχνουν σε αυτόν [3]. Αρχικά σε κάθε κόμβο ανατίθεται ένα score, το PageRank του, και κάθε κόμβος κατανέμει ισόποσα το PageRank του στους γείτονες του κατά μήκος των εξερχόμενων ακμών του. Χωρίς να μπούμε σε λεπτομέρειες, ο αλγόριθμος PageRank αντιστοιχεί σε έναν τυχαίο περιπατητή ο οποίος μεταβαίνει απο κόμβο σε κόμβο με πιθανότητες μετάβασης ανάλογες των PageRank score των γειτόνων του κόμβου στον οποίο βρίσκεται κάθε στιγμή και σαν τον HITS κάνει χρήση πληροφοριάς που περιέχεται στο κυρίαρχο ιδιοδιανύσμα (του 'Google' μητρώου), ενώ όπως και οι τεχνικές Katz και αυτός ο αλγόριθμος απαιτεί την ρύθμιση μιας παραμέτρου (damping factor).

Ενώ η έννοια της σημαντικότητας των κόμβων στον αλγόριθμο PageRank είναι μονοδιάστατη, ο αλγόριθμος HITS στον οποίον θα εστιάσουμε, υποθέτει δύο ειδών σημαντικούς κόμβους σύμφωνα με του ρόλους που μπορεί να έχει κάθε κόμβος ενός κατευθυνόμενου δικτύου, όπως αναφέραμε πιο πάνω. Για παράδειγμα υποθέτωντας ένα δίκτυο που μοντελοποιεί ένα σύνολο εγγράφων, όπως ο Παγκόσμιος Ιστός, κάποιοι κόμβοι είναι πολύτιμοι επειδή παρέχουν σημαντική πληροφορία για ένα συγκεκριμένο θέμα. Αυτοί οι κόμβοι καλούνται authorities. Απο την αλλή κάποιοι άλλοι κόμβοι είναι πολύτιμοι όχι επειδή παρέχουν πληροφορία για κάποιο θέμα αλλά επειδή περιέχουν πληροφορία για το ποιοί κόμβοι περιέχουν την απαιτούμενη πληροφορία. Δείχνουν με άλλα λόγια σε αυτούς τους κόμβους που περιέχουν τη σημαντική πληροφορία. Αυτοί οι κόμβοι καλούνται hubs. Ο αλγόριθμος HITS ορίζει έναν αναδρομικό τρόπο ορισμού αυτών των ποσοτήτων. Καλά hubs είναι εκείνοι οι κόμβοι που δείχνουν σε πολλά καλά authorities, ενώ αντίστοιχα τα καλά authorities είναι εκείνοι οι κόμβοι που έχουν εισερχόμενες ακμές απο πολλά καλά hubs. Συμβολίζοντας με E το σύνολο των ακμών του κατευθυνόμενου γραφήματος και V το σύνολο των κορυφών, ο αλγόριθμος έγκειται σε ένα επαναληπτικό σχήμα κατάταξης στο οποίο σε κάθε κόμβο του δικτύου ανατίθενται δύο βάρη, ένα authority βάρος x_i και ένα hub βάρος y_i, αρχικά ίσα με 1. Σε κάθε επανάληψη k τα βάρη αυτά ενημερώνονται σύμφωνα με τις σχέσεις

$$x_i^{(k)} = \sum_{j:(j,i)\in E} y_j^{(k-1)} \text{ και } y_i^{(k)} = \sum_{j:(i,j)\in E} x_j^{(k-1)} \text{ για k=1,2,3...}$$

Ενώ αν με A συμβολίσουμε τώρα το μητρώο γειτνίασης του κατευθυνόμενου γραφήματος οι παραπάνω σχέσεις μπορούν να γραφούν ως εξής

$$x^{(k)} = A^T y^{(k-1)} \text{ και } y^{(k)} = Ax^{(k)}$$

Αντικαθιστώντας τα $x^{(k)}$ και $y^{(k)}$ τελικά προκύπτουν οι σχέσεις:

$$x^{(k)} = A^T Ax^{(k-1)} \text{ και } y^{(k)} = AA^T y^{(k-1)}$$

Συνήθως τα παραπάνω score κανονικοποιούνται ώστε το άθροισμα των τετραγώνων του κάθενος για το σύνολο των κόμβων να είναι ίσο με 1 ενώ είναι γνωστό πως υπό χαλαρές συνθήκες (Θεώρημα Perron-Frobenius) οι ακολουθίες των διανυσμάτων $x^{(k)}$ και $y^{(k)}$ συγκλίνουν.

Απο την τελευταία σχέση φαίνεται πως ο αλγόριθμος HITS είναι στην ουσία μια μέθοδος δύναμης η οποία υπολογίζει τα κυρίαρχα ιδιοδιανύσματα των μητρώων $A^T A$ και AA^T παράγοντας τα authority και hub scores αντίστοιχα.

Χρησιμοποιώντας την SVD του μητρώου γειτνίασης A προκύπτει πως $AA^T = U\Sigma^2 U^T = U_r \Sigma^2 U_r$ και $A^T A = V\Sigma^2 V^T = V_r \Sigma^2 V_r$, όπου r=rank(A). Επομένως το διάνυσμα που περιέχει τα hub scores είναι το u_1 ενώ το διάνυσμα με τα authority scores το v_1. Επομένως ένας εναλλακτικός τρόπος να δούμε τα hub και authority score είναι σαν τα κυρίαρχα αριστερά και δεξιά ιδιάζοντα διανύσματα του μητρώου A, αντίστοιχα.

3 Εργασία

Αναφέραμε πως οι συγγραφείς επεκτείνουν τα μέτρα κεντρικότητας και επικοινωνίας μεταξύ των κόμβων στα κατευθυνόμενα δίκτυα με σκοπό την εύρεση hubs και authorities αντιστοιχίζοντας το κατευθυνόμενο δίκτυο σε ένα διμερές γράφο με το διπλάσιο πληθος κόμβων. Το πετυχαίνουν αυτό ορίζοντας το επαυξημένο μητρώο

$$\mathcal{A} = \begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix}$$

το οποίο αντιστοιχεί στο μητρώο γειτνίασης του διμερούς μη-κατευθυνόμενου γραφήματος 2n κόμβων, το οποίο προκύπτει απο το αρχικό κατευθυνόμενο δίκτυο. Αυτό το δίκτυο περιέχει 2 αντίγραφα για κάθε κόμβο του αρχικού δικτύου, και αποτελείται από δύο σύνολα κόμβων V και V', όπου υπάρχει μια μη-κατευθυνόμενη ακμή μεταξύ δύο κόμβων των 2 σύνολων αν υπάρχει η αντίστοιχη κατευθυνόμενη ακμή στο αρχικό δίκτυο. Σε αυτό το διμερές γράφο οι n κόμβοι στο V είναι οι κόμβοι του αρχικού κατευθυνόμεου γραφήματος όταν έχουν το ρόλο των πομπών πληροφορίας (δηλαδή τον ρόλο των hubs), ενώ οι n κόμβοι στο σύνολο V' αντιστοιχούν στους ίδιους κόμβους όταν έχουν τον ρόλο των δεκτών πληροφορίας (δηλαδή τον ρόλο των authorities). Σημειώνεται επίσης πως ενώ το μητρώο γειτνίασης ενός κατευθυνόμενου γράφου δεν είναι συμμετρικό και μπορεί να έχει μιγαδικές ιδιοτιμές, πλέον το επαυξημένο μητρώο γειτνίασης $\mathcal{A}$ του διμερούς μη-κατευθυνόμενου δικτύου είναι συμμετρικό με πραγματικές ιδιοτιμές.

Εν συνεχεία οι συγγραφείς παρέχουν έναν εναλλακτικό τρόπο υλοποίησης του αλγορίθμου HITS, εφαρμόζοντας τον αλγόριθμο στο μητρώο $\mathcal{A}$ ορίζοντας το διάνυσμα $u^{(k)} = \begin{pmatrix} y^{(k)} \\ x^{(k)} \end{pmatrix}$ και εκμεταλλεύονται την παραπάνω ιδιότητα συμμετρίας του μητρώου $\mathcal{A}$, οπότε $\mathcal{A}^T = \mathcal{A}$ και $\mathcal{A}^T\mathcal{A} = \mathcal{A}\mathcal{A}^T = A^2$ για να προκύψει η σχέση:

$$u^{(k)} = \mathcal{A}^2 u^{(k-1)} = \begin{pmatrix} AA^T & 0 \\ 0 & A^TA \end{pmatrix} u^{(k-1)}$$

Τα πρώτα n στοιχεία του διανύσματος $u^{(k)}$ αντιστοιχούν στα hub scores ενώ τα τελευταία n στοιχεία στα authority scores. Η σημαντική παρατήρηση εδώ είναι πως ο αλγόριθμος HITS εκμεταλεύεται πληροφορία που περιέχεται μόνο στο κυρίαρχο ιδιοδιάνυσμα του μητρώου $\mathcal{A}$. Όπως θα δούμε στην συνέχεια η επέκταση που προτείνεται απο τους συγγραφείς υπερτερεί του αλγορίθμου HITS ακριβώς λόγω του ότι κάνει χρήση πληροφορίας απο το σύνολο των ιδιοδιανυσμάτων του μητρώου $\mathcal{A}$.

Παρόλο που η χρήση του εκθετικού ενός μητρώου για την εξαγωγή μέτρων κεντρικότητας και επικοινωνίας προσφέρει διαισθητικές και σαφής ερμηνείες στα μη-κατευθυνόμενα γραφήματα, η χρήση του στα κατευθυνόμενα μπορεί να είναι προβληματική. Αιτία γι΄ αυτό αποτελεί το γεγονός πως πλέον στα κατευθυνόμενα δίκτυα δεν είναι δυνατή η ερμηνεία των διαγώνιων στοιχείων του e^A με την χρήση κλειστών περιπάτων καθώς και προβλήματα υπολογιστικού κόστους που μπορεί να προκύψουν. Η ερμηνεία των μη-διαγώνιων στοιχείων του μητρώου e^A ωστόσο παραμένει ίδια και στην περίπτωση κατευθυνόμενων δικτύων. Εξ' αιτίας αυτής της προβληματικής συμπεριφοράς οι συγγραφείς πρότειναν την επέκταση που ακολουθεί.

Με σκοπό να εκμεταλευτούν την συμμετρία που παρέχει το μητρώο γειτνίασης του μη-κατευθυνόμεου διμερούς γραφήματος που προκύπτει απο το κατευθυνόμενο γράφημα, η βασική ιδέα είναι να υπολογιστεί το εκθετικό του μητρώου $\mathcal{A}$ και να ερμηνευτούν κατάλληλα τα στοιχεία του. Χρησιμοποιώντας την SVD για το μητρώο γειτνίασης του κατευθυνόμενου δικτύου, $A = U\Sigma V^T$ παρατηρείται πως το μητρώο γειτνίασης του διμερούς γραφήματος μπορεί να διασπαστεί ως εξής:

$$\mathcal{A} = \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix} \begin{pmatrix} 0 & \Sigma \\ \Sigma & 0 \end{pmatrix} \begin{pmatrix} U^T & 0 \\ 0 & V^T \end{pmatrix}$$

Τελικά, όπως έιχε παρατηρηθεί και στην εργασία [4], το εκθετικό αυτού του μητρώου προκύπτει ίσο με

$$e^{\mathcal{A}} = \begin{pmatrix} \cosh(\sqrt{AA^T}) & A(\sqrt{A^TA})^\dagger \sinh\sqrt{A^TA} \\ \sinh(\sqrt{A^TA})(\sqrt{A^TA})^\dagger A^T & \cosh(\sqrt{A^TA}) \end{pmatrix}$$

3.1 Ερμηνεία Μητρώου

Για την ερμηνεία των στοιχείων αυτού του μητρώου οι ερευνητές λειτουργούν με την εξής λογική. Εάν ένας κόμβος i στο κατευθυνόμενο δίκτυο είναι καλό hub αυτό σημαίνει πως θα υπάρχουν πολλοί εναλλασόμενοι ζυγού μήκους περίπατοι

που εκινούν απο τον κόμβο i, με εξερχόμενη ακμή, και καταλήγουν στον κόμβο i. Το γεγονός αυτό προκύπτει απο τον τρόπο με τον οποίο ορίστηκαν τα καλά hubs σε ένα κατευθυνόμενο δίκτυο. Ένας τρόπος υπολογισμού αυτών των περιπάτων δίνεται χρησιμοποιώντας απο την θεωρία γράφων το γεγονός πως τα στοιχεία $[(AA^T)^k]_{ij}$ και $[(A^TA)^k]_{ij}$ μετράνε το πλήθος των εναλλασόμενων περιπάτων μήκους 2k απο τον κόμβο i στον j. Ζυγίζοντας τώρα κάθε περίπατο μήκους 2k με την ποσότηα $\frac{1}{(2k)!}$ οι ζητούμενοι περίπατοι υπολογίζονται απο το (i,i) στοιχείο του μητρώου

$$I + AA^T + \frac{AA^T}{2!} + \frac{AA^TAA^A}{4!} + ... + \frac{(AA^T)^k}{(2k)!}$$

το οποίο με χρήση της SVD του A ισούται με

$$U\cosh(\Sigma)U^T = \cosh\left(\sqrt{AA^T}\right)$$

Επομένως τα διαγώνια στοιχεία $[e^{\mathcal{A}}]_{ii} = [\cosh\left(\sqrt{AA^T}\right)]_{ii}$ μετράνε την hub κεντρικότητα του κόμβου i, πόσο καλά δηλαδή ο κόμβος i σαν hub μεταδίδει πληροφορία στους authority κόμβους.

Με ανάλογο σκεπτικό εάν ένας κόμβος i είναι καλό authority θα υπάρχουν πολλοί εναλλασόμενοι περιττού μήκους περίπατοι που εκκινούν απο τον κόμβο i, με εισερχόμενη ακμή, και καταλήγουν στον i. Αυτοί οι περίπατοι αντιστοιχούν στο (i,i) στοιχείο του μητρώου $\cosh\left(\sqrt{A^TA}\right)$. Επόμενως δείκτη για το πόσο καλά δέχεται πληροφορία ένας κόμβος i του δικτύου απο τους hub κόμβους αποτελεί η authority κεντρικότητα του κόμβου η οποία δίνεται απο το στοιχείο

$$[e^{\mathcal{A}}]_{n+i,n+i} = [\cosh\left(\sqrt{A^TA}\right)]_{ii}$$

Όσον αφορά τα μη-διαγώνια στοιχεία του πρώτου μπλοκ της διαγωνίου του μητρώου $e^{\mathcal{A}}$, δηλαδή τα στοιχεία $[e^{\mathcal{A}}]_{ij} = [\cosh\left(\sqrt{AA^T}\right)]_{ij}$, αυτά αποτελούν μέτρο της hub επικοινωνίας μεταξύ των κόμβων i και j. Η hub επικοινωνία μεταξύ δύο κόμβων μετράει πόσο όμοιοι είναι δύο κόμβοι σαν hubs, ή αλλιώς το κατα πόσο δείχνουν σε πολλά κοινά authorities του ίδιου θέματος. Αντίστοιχα τα μη-διαγώνια στοιχεία $[e^{\mathcal{A}}]_{n+i,n+j} = [\cosh\left(\sqrt{A^TA}\right)]_{i,j}$ του κάτω δεξιά μπλοκ, αντιστοιχούν στην authority επικοινωνία μεταξύ δύο κόμβων, δηλαδή στο κατά πόσο μοιάζουν στο ρόλο τους σαν authorities. Μεγάλη τιμή authority επικοινωνίας μεταξύ δύο κόμβων δηλώνει πως 'δείχνονται' απο πολλά κοινά hubs και κατα πάσα πιθανότητα περιέχουν παρεμφερή πληροφορία. Τέλος τα στοιχεία των εκτός διαγωνίου μπλοκ του μητρώου $e^{\mathcal{A}}$ αντιστοιχούν στην hub-authority επικοινωνία. Για παράδειγμα υψηλή τιμή στο στοιχείο

$$[e^{\mathcal{A}}]_{n+i,j} = [A(\sqrt{A^TA})^\dagger \sinh\sqrt{A^TA}]_{ij} = [\sinh\left(\sqrt{A^TA}\right)(\sqrt{A^TA})^\dagger A^T]_{ij} = [e^{\mathcal{A}}]_{n+j,i}$$

δηλώνει πως ο κόμβος i τείνει να δείχνει περισσότερο σε κόμβους οι οποίοι περιέχουν πληφορορία όμοια με αυτήν που περιέχει ο κόμβος j σαν authority.

Έχοντας ορίσει τώρα τον νέο τρόπο υπολογισμού hubs και authorities οι συγγραφείς σημειώνουν πως η κατάταξη που επιστρέφει ο αλγόριθμος HITS αποτελεί

προσέγγιση της κατάταξης που επιστρέφεται με την χρήση των διαγώνιων στοιχείων του μητρώου $e^{\mathcal{A}}$ καθώς στη δεύτερη περίπτωση γίνεται χρήση όλων των ιδιοδιανυσμάτων του μητρώου $\mathcal{A}$, σε αντίθεση με τον αλγόριθμο HITS που όπως έχουμε ήδη αναφέρει, χρησιμοποιεί πληροφορία μόνο απο το κυρίαρχο ιδιοδιάνυσμα του μητρώου $\mathcal{A}$. Στην συνέχεια παρατίθενται τα βασικα στοιχεία της απόδειξης τους.

Αν $\lambda_1 > \lambda_2 > ...\lambda_{2n}$ οι ιδιοτιμές και $u_1 > u_2 > ...u_{2n}$ τα κανονικοποιήμενα ιδιοδιανύσματα του μητρώου $\mathcal{A}$ τότε το εκθετικό του μητρώου μπορεί να γραφτεί

$$e^{\mathcal{A}} = e^{\lambda_1} u_1 u_1^T + \sum_{i=2}^{2n} e^{\lambda_i} u_i u_i^T$$

ενώ λαμβάνοντας τα διαγώνια στοιχεία που μας ενδιαφέρουν και κλιμακώνοντας κατάλληλα με την ποσότητα e^{λ_1} έχουμε

$$diag(e^{-\lambda_1} e^{\mathcal{A}}) = diag(e^{\mathcal{A}-\lambda_1 I}) = diag(u_1 u_1^T) + \sum_{i=2}^{2n} e^{\lambda_i - \lambda_1} diag(u_1 u_1^T)$$

Οπότε η χρήση των διαγώνιων στοιχείων του εκθετικού του μητρώου $\mathcal{A}$ για την ταξινόμηση hubs και authorities συνίσταται στην εύρεση των τετραγώνων όλων των στοιχείων των ιδιοδιανυσμάτων του $\mathcal{A}$, ζυγισμένα απο το εκθετικό της αντίστοιχης ιδιοτιμής. Απο την άλλη, όπως φαίνεται απο την παραπάνω σχέση, ο αλγόριθμος HITS συνίσταται στην χρήση μόνο του κυρίαρχου όρου στο παραπάνω ανάπτυγμα, δηλαδή στην προσέγγιση $e^{\mathcal{A}} \approx e^{\lambda_1} u_1 u_1^T$. Προφανώς αν η φασματική διαφορά (spectral gap) είναι μεγάλη ($\lambda_1 >> \lambda_2$), τότε τα αποτελέσματα του HITS δεν αναμένεται να διαφέρουν πολύ απο αυτά που λαμβάνουμε με την χρήση των διαγώνιων στοιχείων του $e^{\mathcal{A}}$ καθώς μεγάλο μέρος της πληφορφορίας περιέχεται στον πρώτο όρο της παραπάνω σχέσης. Αντίθετα αν $\lambda_1 \approx \lambda_2$ τότε η συνεισφορά απο τα υπόλοιπα ιδιοδιανύσματα πέραν του κυρίαρχου, που αντιστοιχεί στο άθροισμα στην παραπάνω σχέση, θα φέρει σημαντική πληροφορία που δεν λαμβάνεται καθόλου υπ΄ όψιν απο τον HITS. Τέλος προφανώς οι δύο προσεγγίσεις που συγκρίνονται εδώ αποτελούν ακραίες περιπτώσεις. Μια ενδιάμεση προσέγγιση, η οποία θα έκανε χρήση ενός υποσυνόλου των ιδιοδιανυσμάτων του μητρώου $\mathcal{A}$, θα μπορούσε να επιφέρει βελτιώσεις στα αποτελέσματα του HITS.

Επιπλέον σχετικά με τα παραπάνω αξίζει να σημειώσουμε εδώ πως οι συγγραφείς δεν αναφέρονται στην τάξη της διαφοράς μεταξύ λ_1 και λ_2 που θεωρείται σημαντική ή αν είναι μόνο το spectral gap που κάνει τη διαφορά. Ίσως θα ήταν δόκιμο να εξετασστεί εμπειρικά ποιάς τάξης διαφορά μεταξύ των δύο μεγαλύτερων ιδιοτιμών επιφέρει σημαντικά διαφορετικά αποτελέσματα μεταξύ των δύο προσσεγίσεων, αν μπορεί να καθοριστεί ένα τέτοιο κατώφλι καθώς και κατά πόσο επηρεάζονται τα αποτελέσματα του αλγορίθμου HITS απο αυτή την διαφορά.

Οι συγγραφείς αναφέρονται συνοπτικά και στις εναλλακτικές επιλογές που υπάρχουν όσον αφορά τις συναρτήσεις, πέραν του εκθετικού, που μπορούν να χρησιμοποιηθούν για την ταξινόμηση των κόμβων σε hubs και authorities στα κατευθυνόμενα δίκτυα. Πιο συγκεκριμένα επεκτείνουν την χρήση της συνάρτησης resolvent εφαρμόζοντας τη στο μητρώο γειτνίασης του διμερούς γραφήματος

$\mathcal{A}$ που προκύπτει απο το κατευθυνόμενο δίκτυο. Αυτό έχει σαν αποτέλεσμα το μητρώο

$$R(A;c) = \begin{pmatrix} (I - c^2 AA^T)^{-1} & cA(I - c^2 A^T A)^{-1} \\ c(I - c^2 A^T A)^{-1} A^T & (I - c^2 A^T A)^{-1} \end{pmatrix}$$

όπου $1 < c < 1/\sigma_1$ και σ_1 η μέγιστη ιδιάζουσα τιμή του μητρώου γειτνίασης του κατευθυνόμενου γράφου A. Παρόμοια με τη χρήση του εκθετικού, τα διαγώνια στοιχεία του μπλοκ $(I - c^2 AA^T)^{-1}$ παρέχουν τα hub scores ενώ τα διαγώνια στοιχεία του μπλοκ $(I - c^2 A^T A)^{-1}$ παρέχουν τα authority scores.

Ακόμη hub και authority score για κάθε κόμβο μπορούν να ανακτηθούν με την κεντρικότητα Katz. Τα authority scores για κάθε κόμβο δίνονται εξ' ορισμού απο το άθροισμα των γραμμών του μητρώου $(I - cA)^{-1}$ ενώ για τα hub scores προκύπτουν απο το άθροισμα των γραμμών του μητρώου $(I - cA^T)^{-1}$, όπου A το μητρώο γειτνίασης του κατευθυνόμενου δικτύου.

Ο τελευταίος αλγόριθμος που εξετάζεται για την εύρεση hubs και authorities είναι ο PageRank. Η εφαρμογή του PageRank έχει σαν αποτέλεσμα την κατάταξη των κόμβων σε authorities. Για την ανάδειξη των κόμβων hubs χρησιμοποιείται μια παραλλαγή, η οποία ονομάζεται Reverse PageRank και συνίσταται στην εφαρμογή του PageRank στο ανάστροφο μητρώο γειτνίασης A^T. Αυτό αντιστοιχεί στην αντιστροφή της ακμής (i,j) στο κατευθυνόμενο δίκτυο σε μία ακμή (j,i). Η παραλλαγή αυτή χρησιμοποιείται συνήθως όταν πρέπει να καθοριστεί γιατί ένας κόμβος είναι σημαντικός αντί για το ποιοί κόμβοι είναι σημαντικοί. Διαισθητικά ο αλγόριθμος Reverse PageRank μοντελοποιεί ένα τυχαίο περιπατητή, ο οποίος ακολουθεί εισερχόμενες ακμές αντί για εξερχόμενες. Έτσι μεγάλες τιμές Reverse PageRank υποδυκνείουν κόμβους που μπορούν να προσεγγίσουν πολλούς κόμβους του δικτύου, με άλλα λόγια καλά hubs. Βέβαια σε αντίθεση με την τεκμηριωμένη αποτελεσματικότητα του αλγορίθμου PageRank για την εύρεση authorites, η αποτελεσματικότητα του αλγορίθμου Reverse PageRank για την εύρευση hubs είναι αμφίβολη καθώς δεν έχει εξεταστεί ιδιαίτερα.

3.2 Πειραματικό Σκέλος

Οι συγγραφείς εξέτασαν πειραματικά τα όσα αναλύσαμε πιο πάνω σε 4 πραγματικά σύνολα γραφημάτων ιστοτόπων διαφορετικής θεματολογίας. Η κατάταξη των κόμβων σε hubs και authorities που προκύπτει απο τα διαγώνια στοιχεία του μητρώου $e^{\mathcal{A}}$ συγκρίνεται με τις αντίστοιχες κατατάξεις που προκύπτουν απο τον αλγόριθμο HITS, την κεντρικότητα Katz με $c = 1/(\rho(A) + 0.1)$, τα αθροίσματα γραμμών και στηλών του εκθετικού του μη-συμμετρικού μητρώου γειτνίασης του κατευθυνόμενου δικτύου e^A και του αλγορίθμου PageRank/Reverse PageRank.

Το πρώτο σύνολο δεδομένων με θέμα τις εκτρώσεις αποτελείται απο 2293 κόμβους. Εδώ οι συγγραφείς παρατηρούν πως οι κατατάξεις που προκύπτουν απο το $e^{\mathcal{A}}$ και τον αλγόριθμο HITS ομοιάζουν αρκετά, με κοινές τις 6 πρώτες hub ιστοσελίδες και τις 7 πρώτες authority ιστοσελίδες, αλλά με διαφορετική σειρά και στις δύο περιπτώσεις. Το αποτέλεσμα αυτό είναι δικαιολογημένο με βάση το σχετικά μεγάλο spectral gap, αφού η μέγιστη και η αμέσως επόμενη ιδιοτιμές είναι ίσες με $\lambda_N = 31.91$ και $\lambda_{N-1} = 26.04$ αντίστοιχα. Επίσης η κατάταξη με τη χρήση του

μητρώου e^A συμπίπτει σε μεγάλο βαθμό με αυτή της μετρικής Katz αλλά και οι δύο διαφέρουν σημαντικά απο τις κατατάξεις του αλγορίθμου HITS και του $e^{\mathcal{A}}$, ενώ η κατάταξη που επιστρέφει ο συνδυασμός των αλγορίθμων PageRank/Reverse PageRank διαφέρει διαφέρει σημαντικά απο όλες τις υπόλοιπες.

Το δεύτερο γράφημα με αντικείμενο την υπολογιστική πολυπλοκότητα αποτελείται απο 884 κόμβους. Εδώ το spectral gap είναι αρκετά μικρότερο απο ότι πριν και ίσο με $\lambda_N - \lambda_{N-1} = 10.93 - 9.86 = 1.07$. Όπως αναμενόταν τα αποτελέσματα μεταξύ HITS και της διαγωνίου του $e^{\mathcal{A}}$ διαφέρουν αρκετά με μόλις 3 κοινά hubs στην πρώτη δεκάδα, με κοινό τον πρώτο κόμβο και 4 κοινούς authority κόμβους με αρκετά διαφορετική κατάταξη. Οι μετρικές Katz και e^A δεν συμβαδίζουν σε αυτήν την περίπτωση ενώ επιβεβαιώνεται εμπειρικά η εξάρτηση της μετρικής Katz απο την παράμετρο c και παρατηρείται πως η χρήση του PageRank/Reverse PageRank έχει παρόμοια αποτελέσματα με την e^A ενώ η λιγότερο αξιόπιστη μέθοδος εδώ είναι ο αλγόριθμος HITS εξ' αιτίας του μικρού spectral gap.

Το τρίτο σύνολο που εξετάστηκε έχει να κάνει με την θανατική ποινή και αποτελείται απο 1850 κόμβους. Εδώ το spectral gap είναι ίσο με $\lambda_N - \lambda_{N-1} = 28.02 - 17.68 = 10.34$, αρκετά μεγαλύτερο απο τις προηγούμενες δύο περιπτώσεις. Αυτό έχει σαν αποτελέσμα οι κατατάξεις που παράγουν ο HITS και τα διαγώνια στοιχεία του $e^{\mathcal{A}}$ να είναι ταυτόσημες. Η χρήση Katz και e^A παράγουν παρεμφερή αποτελέσματα αλλά διαφορετικά απο αυτά των HITS και $e^{\mathcal{A}}$, ενώ τα αποτελέσματα του αλγορίθμου PageRank/Reverse PageRank δεν συμπίπτουν με αυτά των προηγούμενων τεχνικών.

Τελός δοκιμάζεται και το γράφημα wb-cs-stanford το οποίο αποτελείται απο 9914 κόμβους. Το spectral gap για το μητρώο $\mathcal{A}$ είναι $\lambda_N - \lambda_{N-1} = 38.38 - 32.12 = 6.26$. Σε αυτήν την περίπτωση παρατηρήθηκε σημαντική επικάλυψη μεταξύ του HITS, της διαγωνίου του $e^{\mathcal{A}}$, της μετρικής Katz και των αθροισμάτων γραμμών/στηλών του e^A, με τα αποτελέσματα μόνο του PageRank/Reverse PageRank να διαφέρουν. Αυτή η συμφωνία μεταξύ των μετρικών εικάζεται ότι οφείλεται στην συμμετρία που έχει το μητρώο γειτνίασης A, σε ποσοστό 47.63 % έναντι 2-4% των προηγουμένων.

3.3 Προσέγγιση Εκθετικού ενός Μητρώου

Οι συγγραφείς τονίζουν πως όταν μας είναι απαραίτητο μόνο ένα υποσύνολο των στοιχείων του εκθετικού ενός πραγματικού συμμετρικού μητρώου A (στην προεκιμένη περίπτωση η διαγώνιος), η οποιασδήποτε άλλης συνάρτησης μητρώου, ή μας ενδιαφέρει περισσότερο η κατάταξη των στοιχείων και όχι οι ακριβείς τιμές τους, είναι απαραίτητες αποδοτικές τεχνικές για την εκτίμηση ανω και κάτω φράγματων των στοιχείων αυτών, έναντι κοστοβόρων τεχνικών για τον ακριβή υπολογισμό όλων των στοιχείων του εκθετικού του μητρώου A, όπως η συνάρτηση expm της MATLAB κόστους $O(n^3)$, αν το μητρώο A είναι διαστάσεων $n \times n$

Δείχνουν ότι ο υπολογισμός των στοιχείων μιάς αυστηρά εντελώς μονότονης (s.c.m) [4] συνάρτησης ενός μητρώου f(A) συνίσταται στον υπολογισμό διγραμμικών εκφράσων της μορφής $u^T f(A)v$, για δοθέντα διανύσματα u και v, ενώ αυτές οι διγραμμικές εκφράσεις μπορούν να αναχθούν στον υπολογισμό Riemann-Stieltjes ολοκληρωμάτων. Αυτά τα ολοκληρώματα προσεγγίζονται με κανόνες αριθμητικής

ολοκλήρωσης Gauss στους οποίους η συνάρτηση βάρους αντιστοιχεί σε συνάρτηση των ιδιοτιμών του μητρώου A.

Είναι γνωστό απο την βιβλιογραφία πως οι κανόνες ολοκληρώσης Gauss για θετικά ορισμένες συναρτήσεις, όπως είναι το εκθετικό και το resolvent που μας ενδιαφέρουν εδώ, είναι ισοδύναμες με τον υπολογισμό των στοιχείων και των ιδιοτιμών/ιδιοδιανυσμάτων του συμμετρικού και τριδιαγώνιου μητρώου Hessenberg που προκύπτει απο τον αλγόριθμο Lanczos. Τα στοιχεία του μητρώου Hessenberg υπολογίζονται εφαρμόζωντας τον αλγόριθμο Lanczos με αρχικά διανύσματα $x_{-1} = 0$ και $x_0 = e_i$, ενώ οι ιδιοτιμές του μητρώου Hessenberg αντιστοιχούν στα σημεία παρεμβολής Gauss και τα στοιχεία των κανονικοποιήμενων ιδιοδιανυσμάτων του μητρώου Hessenberg, στο τετράγωνο, παρέχουν τις τιμές της συνάρτησης βάρους. Εναλλακτικά μπορούμε να δούμε τη παραπάνω διαδικασία υπολογισμού των κανόνων ολοκλήρωσης Gauss ως τον υπολογισμό μιας ακολουθίας ορθογώνιων, ως προ το εσωτερικό γινόμενο, πολυωνύμων μέσω μιας αναδρομής τριών όρων.

Τέλος για το πρόβλημα της εύρεσης των διαγώνιων στοιχείων του μητρώου $e^{\mathcal{A}}$, οι συγγραφείς επιβεβαιώνουν πειραματικά πως η υπολογιστική πολυπλοκότητα χρήσης των μεθόδων ολοκλήρωσης Gauss για την εύρεση των score των κόμβων είναι της τάξης του O(n), για ένα κόμβο, και O(n^2) συνολικά, ενώ το πλήθος των επαναλήψεων Lanczos που απαιτούνται εξαρτάται απο τις ιδιοτιμές του μητρώου $e^{\mathcal{A}}$ και τη διαφορά στις τιμές των διαγώνιων στοιχείων του. Πράγματι για τα σύνολα δεδομένων που εξετάστηκαν παραπάνω για την κατάταξη των κόμβων, στις περισσότερες περιπτώσεις απαιτούνται 2 έως 5 επαναλήψεις Lanczos ανά κόμβο, με μόνη εξαίρεση το γράφημα των εκτρώσεων στο οποίο για το διαχωρισμό μεταξύ των πρώτων 10 hubs απαιτούνται πάνω απο 40 επαναλήψεις λόγω του ότι τα scores αυτών των κόμβων συμφωνούν σε 15 δεκαδικά ψηφία.

4 Δικά μου Πειράματα και Συμπεράσματα

Σε αυτήν την ενότητα παραθέτω ορισμένα αποτελέσματα και παρατηρήσεις σχετικά με κάποια πειράματα που εκτέλεσα και αφορούν την σύγκριση μεταξύ των κατατάξεων των κόμβων σε hubs και authorities που προκύπτουν απο τον αλγόριθμο HITS και τα διαγώνια στοιχεία του μητρώου $e^{\mathcal{A}}$.

Αρχικός στόχος των πειραμάτων μου ήταν να εξερευνήσω ποια διαφορά στο spectral gap θεωρείται σημαντική ώστε η κατάταξη που επιστρέφουν οι δύο παραπάνω τεχνικές να αρχίσουν να αποκλίνουν σημαντικά ή αν και κατά πόσο μπορεί να παίζουν ρόλο και άλλοι παράγοντες. Η συμπεριφορά των περισσότερων γραφημάτων που δοκίμασα συνάδει απόλυτα με τα όσα αναλύθηκαν στην εργασία. Εδώ θα παρουσιάσω κάποιες ειδικές περιπτώσεις γραφημάτων στις οποίες παρατηρήθηκαν ΄περίεργες΄ θα μπορούσε να πεί κανείς συμπεριφορές που δεν συνάδουν απόλυτα με τα όσα αναλύθηκαν στην εργασία [1].

Αυτό που έκανα ήταν να ορίσω δύο συναρτήσεις οι οποίες λαμβάνοντας το μητρώο γειτνίασης ενός κατευθυνόμενου δικτύου επιστρέφουν τους κορυφαίους 10 hub και authority κόμβους αντίστοιχα, που αναγνωρίζει ο αλγόριθμος HITS. Επίσης όρισα άλλες δύο συνάρτησεις οι οποίες δέχονται σαν όρισμα το εκθετικό του επαυξημένου μητρώου γειτνίασης $e^{\mathcal{A}}$ και επιστρέφουν τους 10 κορυφαίους

hub και authority κόμβους με τον ίδιο τρόπο ακριβώς που υποδεικνύεται στην εργασία. Ο κώδικας, μαζί με επιπλέον στοιχεία για κάθε γράφημα που δεν έχουν συμπεριληφθεί στην εργασία, είναι διαθέσιμος εδώ. Όλα τα γραφήματα που έχω χρησιμοποιήσει είναι απο την συλλογή SuiteSparse Matrix Collection.

Όλα τα πειράματα υλοποιήθηκαν με τη χρήση της MATLAB 2016a σε ένα λάπτοπ TOSHIBA Quosmio x770-10N με επεξεργαστή Intel® Core™ i7-2630QM 2.00/2.90 GHz και 8 GB RAM.

Για τα παρακάτω όταν αναφερόμαστε σε ιδιοτιμές και spectral gap μιλάμε πάντα για το επαυξημένο μητρώο $\mathcal{A}$ διαστάσεων $N \times N$ (N=2n) που έχει κατασκευαστεί και αντιστοιχεί στο διμερές μη-κατευθυνόμενο γράφο που έχει προκύψει απο το κατευθυνόμενο δίκτυο με το οποίο ασχολούμαστε σε κάθε περίπτωση.

Αρχικά όσον αφορά το γράφημα Harvard 500 της Mathworks, το αντίστοιχο μητρώο $\mathcal{A}$ έχει πολύ μικρό spectral gap ίσο με 0.0448 καθώς οι 2 μεγαλύτερες ιδιοτιμές του είναι ίσες με $\lambda_N = 18.1480$ και $\lambda_{N-1} = 17.7$. Παρόλα αυτά όπως φαίνεται και από τους πίνακες 1 και 2 ενώ η κατάταξη των κόμβων, με τις δύο τεχνικές, σε authorities έχει εντελώς διαφορετικά αποτελέσματα, οι δύο κατατάξεις των κόμβων σε hubs παρουσιάζουν μεγάλη ομοιότητα αντίθετα απο ότι θα περιμέναμε για μία τόση μικρή τιμή spectral gap.

$[e^{\mathcal{A}}]_{ii}$	HITS
1	1
231	229
234	231
236	232
238	234
239	236
240	237
229	238
232	239
237	240

Πίνακας 1: Γράφημα Harvard500: Top 10 hubs

$[e^{\mathcal{A}}]_{ii}$	HITS
316	235
321	229
322	230
324	231
325	232
329	233
333	236
318	237
319	238
320	240

Πίνακας 2: Γράφημα Harvard500: Top 10 authorities

Παρόμοια συμπεριφορά παρατηρείται και στο γράφημα Pajek/GD95b (73 κόμβους και 96 ακμές) για το οποίο προκύπτει spectral gap ίσο με $\lambda_N - \lambda_{N-1} = 4.7938 - 4.3655 = 0.4283$, αλλά οι πρώτοι 7 hubs προκύπτουν ίδιοι, με εξαίρεση τον δεύτερο κόμβο, ενώ και οι πρώτοι 5 authority κόμβοι είναι ίδιοι. Χάριν συντομίας δεν παρατίθενται οι σχετικούς πίνακες.

Για το γράφημα CollegeMsg απο την συλλογή SNAP το οποίο αποτελείται απο 1899 κόμβους και 20296 ακμές παρατηρούμε απο τους πίνακες 3 και 4 πως με τις δύο διαφορετικές μεθόδους προκύπτουν μόνο 2 κοινά hubs, αλλά σε διαφορετική σειρά, με όλα τα άλλα να διαφέρουν και μόλις 1 κοινός κόμβος authority, με όλους τους άλλους να είναι εντελώς διαφορετικοί στην πρώτη δεκάδα.

Δεδομένου πως το spectral gap για το συγκεκριμένο γράφημα είναι ίσο με

$[e^{\mathcal{A}}]_{ii}$	HITS
12	103
9	9
323	105
105	400
398	249
1624	32
431	41
372	357
103	12
1168	67

Πίνακας 3: SNAP/CollegeMsg: Top 10 hubs

$[e^{\mathcal{A}}]_{ii}$	HITS
569	598
118	32
1312	638
1624	840
8	713
282	475
711	72
32	97
298	626
341	502

Πίνακας 4: SNAP/CollegeMsg: Top 10 authorities

$\lambda_N - \lambda_{N-1} = 229.350 - 213.8255 = 15,5245$ θα περιμέναμε τα αποτελέσματα του HITS να είναι αρκετά κοντά σε αυτά που προκύπτουν απο το $e^{\mathcal{A}}$. Αντίθετα εδώ βλέπουμε πως τα αποτέλεσματα αποκλίνουν σημαντικά απο το αναμενόμενο.

Αντίθετα για το γράφημα Email-Eu-core της συλλογής SNAP, το οποίο αποτελείται απο 1005 κόμβους και 25571 ακμές, οι δύο τεχνικές παράγουν ακριβώς τα ίδια αποτελέσματα όσον αφορά τις πρώτες 10-άδες αν και εδώ έχουμε σημαντικά μεγαλύτερο spectral gap ίσο με $\lambda_N - \lambda_{N-1} = 64.9012 - 33.2997 = 31.6015$. Επειδή τα αποτελέσματα, όσον αφορά τις πρώτες 10-άδες κόμβων, είναι ταυτόσημα γι αυτό το γράφημα δεν παραθέτουμε τους αντίστοιχους πίνακες.

Ένα άλλο γράφημα που εξετάστηκε είναι το Roget απο την συλλογή Pajek με 1022 κόμβους και 5075 ακμές. Τα αποτελέσματα γι αυτό το γράφο φαίνονται στους παρακάτω πίνακες 5 και 6. Οι 2 μεγαλύτερες ιδιοτιμές του αντίστοιχου μητρώου A είναι $\lambda_N = 9.068$ και $\lambda_{N-1} = 7.8624$, οπότε το spectral gap είναι ίσο με 1.3244.

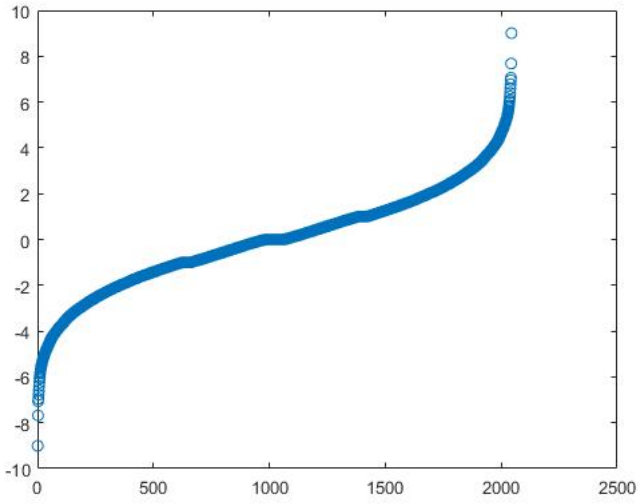
$[e^{\mathcal{A}}]_{ii}$	HITS
664	714
507	507
539	664
714	511
511	539
540	540
674	470
660	713
721	660
688	688

Πίνακας 5: Pajek/Roget: Top 10 hubs

$[e^{\mathcal{A}}]_{ii}$	HITS
557	557
660	660
556	470
698	556
470	698
539	507
674	469
469	674
562	539
507	486

Πίνακας 6: Pajek/Roget: Top 10 authorities

Εδώ παρά το σχετικά μικρό spectral gap βλέπουμε μια σχετική συμφωνία μεταξύ των δύο μεθόδων. Ως προς την κατάταξη των hub κόμβων προκύπτουν 8 κοινοί μεταξύ των δύο μεθόδων αλλά με διαφορετική σειρά. Ως προς τους authority κόμβους και οι δύο μέθοδοι αναγνωρίζουν 8 κοινούς κόμβους στην πρώτη 10-άδα με ίδιους τους κορυφαίους 2. Αυτό που κάνει ακόμα πιο περίεργη αυτήν την σχετική συμφωνία μεταξύ των δύο μεθόδων για το παρόν γράφημα είναι η κατανομή των ιδιοτιμών του μητρώου $\mathcal{A}$.



Σχήμα 1: Ιδιοτιμές του επαυξημένου μητρώου $\mathcal{A}$ για το γράφημα Pajek/Roget

Όπως φαίνεται στην παραπάνω εικόνα 1, μπορεί η πρώτη με την δεύτερη ιδιοτιμή να είναι σχετικά καλά διαχωρισμένες αλλά απο την δεύτερη ιδιοτιμή και μετά οι απόλυτες διαφορές μεταξύ των ιδιοτιμών είναι πάρα πολύ μικρές. Αυτό μου λέει πως η πληροφορία που φέρει το κυρίαρχο ιδιοδιάνυσμα του μητρώου $\mathcal{A}$ και εκμεταλεύεται ο αλγόριθμος HITS θα είναι περιορισμένη. Σημειώνεται ότι το παρών γράφημα αποτελείται απο 77 ισχυρές συνεκτικές συνιστώσες και 21 ασθενώς συνεκτικές συνιστώσες.

Συγκρίσιμο spectral gap με τη προηγούμενη περίπτωση, για το αντίστοιχο μητρώο $\mathcal{A}$, προκύπτει από το γράφημα CSPhd και πάλι απο την συλλογή Pajek, το οποίο αποτελείται απο 1882 κόμβους και 1740 ακμές. Εδώ έχουμε $\lambda_N - \lambda_{N-1} = 6.7099 - 4.8008 = 1.9091$ αλλά τώρα όπως φαίνεται απο τους πίνακες 7 και 8 και ο αλγόριθμος HITS και η χρήση των διαγώνιων στοιχείων του $e^{\mathcal{A}}$ αναγνωρίζουν τον ίδιο κόμβο σαν το πιο σημαντικό hub αλλά διαφέρουν σε όλους τους υπόλοιπους. Το ίδιο παρατηρείται και για τους κόμβους authorities. Οι δύο τεχνικές επιστρέφουν τον ίδιο κορυφαίο κόμβο αλλά δεν έχουν κανένα άλλον κοινό μεταξύ των πρώτων 10.

Οξύμωρο το γεγονός πως στην μία περίπτωση φασματική διαφορά ίση με 1.32

$[e^{\mathcal{A}}]_{ii}$	HITS
216	216
18	463
132	66
20	367
258	427
196	1638
32	826
461	56
207	357
826	10

Πίνακας 7: Pajek/CSPhd: Top 10 hubs

$[e^{\mathcal{A}}]_{ii}$	HITS
1637	1637
744	149
611	215
786	378
1061	448
1303	468
378	493
448	522
522	572
639	611

Πίνακας 8: Pajek/CSPhd: Top 10 authorities

φαίνεται ικανή ώστε τα αποτελέσματα του HITS να αρχίζουν να προσσεγγίζουν, έστω και με κάποιο σφάλμα, τα αποτελέσματα των στοιχείων $[e^{\mathcal{A}}]_{ii}$ ενώ στην άλλη λίγο μεγαλύτερη φασματική διαφορά ίση με 1.9091 φαίνεται πολύ μικρή καθώς μόνο οι 2 πρώτοι κόμβοι hub και authority είναι ταυτόσημοι μεταξύ των δύο τεχνικών, που σημαίνει πως εκτός του κυρίαρχου ιδιοδιανύσματος υπάρχει σημαντική πληροφορία που δεν λαμβάνει υπ'όψιν του ο αλγόριθμος HITS.

Τέλος το γράφημα GD095c και πάλι απο τη συλλογή Pajek το οποίο αποτελείται από 62 κόμβους και 287 ακμές έχει σαν αποτελέσμα μητρώο $\mathcal{A}$ το οποίο παρουσιάζει spectral gap ίσο με $\lambda_N - \lambda_{N-1} = 6.4805 - 5.1894 = 1.2911$. Όπως φαίνεται απο τους πίνακες 9 και 10 εδώ spectral gap ίσο με 1.2911 δείχνει να είναι ικανό ώστε τα αποτελέσματα του HITS να συγκλίνουν σε αυτά που προκύπτουν απο τα διαγώνια στοιχεία του $e^{\mathcal{A}}$.

$[e^{\mathcal{A}}]_{ii}$	HITS
1	1
3	3
6	6
2	2
5	5
28	4
4	28
35	41
30	11
29	9

Πίνακας 9: Pajek/GD95c: Top 10 hubs

$[e^{\mathcal{A}}]_{ii}$	HITS
1	1
3	3
6	6
2	2
5	5
28	4
4	28
35	41
30	11
29	9

Πίνακας 10: Pajek/GD95c: Top 10 authorities

Σημειώνεται πως οι κόμβοι που αναγνωρίζονται απο τα στοιχεία $[e^{\mathcal{A}}]_{ii}$ σαν

κορυφαία hubs είναι ακριβώς οι ίδιοι με αυτούς που αναγνωρίζονται σαν κορυφαία authorities. Αυτό είναι αναμενόμενο εδώ και οφείλεται στο γεγονός πως το μητρώο γειτνίασης του γραφήματος Pajek/GD095c παρουσίαζει 99.7% συμμετρία.

Συμπερασματικά με βάση τα παραπάνω πειράματα έχω να παρατηρήσω τα εξής. Πρώτα από όλα είναι δυνατόν οι διαφορές μεταξύ του αλγορίθμου HITS και των διαγώνιων στοιχείων του $e^{\mathcal{A}}$ να εντοπίζονται μόνο στο ένα υποσύνολο κόμβων, δηλαδή είτε οι κόμβοι hubs να είναι εντελώς διαφορετικοί αλλά τα authorities να παρουσιάζουν ομοιότητες είτα το αντίστροφο, όπως φαίνεται στο γράφημα Harvard500. Πάνω σε αυτό το θέμα δεν γίνεται σαφές, ούτε απο την εργασία, αν και κατά πόσο αυτή η συμπεριφορά, το να προκύπτει δηλαδή κατάταξη πολύ κοντά στην πραγματική απο τον αλγόριθμο HITS αλλά μόνο για το ένα υποσύνολο κόμβων, μπορεί να θεωρηθεί ως αποτελέσμα της, έστω και ελάχιστης, πληροφορίας που λαμβάνεται απο το κυριάρχο ιδιοδιάνυσμα του μητρώου $\mathcal{A}$ ή απορρέει απο κάποια άλλο δομικό χαρακτηριστικό των γραφημάτων το οποίο ευνοεί την ανίχνευση μόνο του ενός είδους κόμβων με σχετική ευκολία (χρησιμοποιώντας μόνο το κυρίαρχο ιδιοδιάνυσμα) έναντι του άλλου.

Εν συνεχεία με βάση τα πειράματα που εκτελέστηκαν καταλήγουμε στο συμπέρασμα πως είναι δύσκολο να καθοριστεί ένα κατώφλι για το spectral gap το οποίο να κάνει τη διαφορά μεταξύ των δύο μεθόδων. Λόγου χάρη για τα γραφήματα Roget και GD95c της συλλογής Pajek βλέπουμε πως με spectral gap 1.3244 και 1.2911 αντίστοιχα λαμβάνουμε και στις δύο περιπτώσεις κατατάξεις των κόμβων σε hubs και authorites με πολλές ομοιότητες μεταξύ των δύο μεθόδων. Και ενώ το γεγονός αυτό θα μπορούσε να αποτελέσει ένδειξη πως μια φασματική διαφορά αυτής της τάξης μπορεί να θεωρηθεί σημαντική ώστε τα αποτελέσματα του HITS να αρχίζουν να συμπίπτουν με αυτά που προκύπτουν απο τα διαγώνια στοιχεία του $e^{\mathcal{A}}$, αντίθετα βλέπουμε πως για το γράφημα CSPhd με spectral gap=1.9091, πλην των πρώτων κόμβων στις κατατάξεις, όλοι οι άλλοι κόμβοι στις πρώτες δεκάδες είναι διαφορετικοί μεταξύ των δύο μεθόδων. Βέβαια όσον αφορά την παραπάνω σύγκριση οφείλουμε να κάνουμε τις ακόλουθες παρατηρήσεις, αν και δεν είμαστε σε θέση να γνωρίζουμε κατά πόσο αυτές μπορεί να επηρεάζουν το πόσο αποτελεσματικά ο HITS προσεγγίζει τα αποτελέσματα των $[e^{\mathcal{A}}]_{ii}$. Πρώτον το μητρώο γειτνίασης του γραφήματος CSPhd είναι εξαιρετικά αραίο, καθώς το γράφημα περιέχει λιγότερες ακμές απο κόμβους σε αντίθεση με τα άλλα 2 γραφήματα. Δεύτερον το μητρώο γειτνίασης του γραφήματος CSPhd δεν παρουσιάζει την παραμικρή συμμετρία σε αντίθεση με τα μητρώα γειτνίασης των Roget και GD95c τα οποία παρουσίαζουν 56.2% και 99.7% συμμετρία, αντίστοιχα.

Μια λογική εξήγηση για την αντίθεση που παρατηρείται στα παραπάνω γραφήματα είναι πως ίσως παίζει ρόλο το σχετικό spectral gap συναρτήσει του μεγέθους των ιδιοτιμών. Για παράδειγμα για το δίκτυο Email-Eu-core η τιμή 31.6015 του spectral gap αντιστοιχεί σχεδόν στο μισό της μεγαλύτερης ιδιοτιμής $\lambda_N = 64.9012$. Αντίθετα για το δίκτυο CollegeMsg παρά το γεγονός πως έχουμε μια μεγάλη τιμή ίση με 15,5245 στο spectral gap, αυτή η τιμή είναι μικρή συγκριτικά με την τάξη μεγέθους των ιδιοτιμών, αντιστοχεί μόλις στο 6.77% της μέγιστης ιδιοτιμής. Δεν θα είχε νόημα επομένως ο καθορισμός μιας τιμής κατωφλίου για το spectral gap πάνω απο την οποία τα αποτελέσματα του αλγορίθμου HITS θα αρχίζουν να συγκλίνουν σε αυτά που προκύπτουν χρησιμοποιώντας το μητρώο

$e^{\mathcal{A}}$, αφού η τιμή αυτή θα εξαρτάται απο την τάξη μεγέθους των ιδιοτιμών του εκάστοτε προβλήματος. Βέβαια δεν είναι σκόπιμο να προβούμε σε συμπεράσματα με βάση μόνο τα δύο αυτά γραφήματα και περισσότερα πειράματα θα ήταν αναγκαία προκειμένου να επιβεβαιωθεί αυτός ο ισχυρισμός.

Σαν τελικό σχόλιο για να αποφανθούμε πότε ο αλγόριθμος HITS έχει αποτελέσματα που προσεγγίζουν με μεγάλη ακρίβεια αυτά των διαγώνιων στοιχείων του $e^{\mathcal{A}}$, ίσως μεγαλύτερη σημασία απο το spectral gap να έχει το πλήθος και το μέγεθος των υπόλοιπων ιδιοτιμών. Πιο συγκεκριμένα αν έχουμε ένα φαινομενικά μεγάλο spectral gap αλλά υπάρχουν πολλές ιδιοτιμές με τιμές κοντά στην τιμή της δεύτερης μεγαλύτερης ιδιοτιμής τότε παρά το μεγάλο spectral gap, τα ιδιοδιανύσματα που αντιστοιχούν σε αυτές τις ιδιοτιμές θα φέρουν σημαντική πληροφορία που δεν εκμεταλεύεται ο HITS. Αν απο την άλλη αν οι ιδιοτιμές είναι πιο εύκολα ΄διαχωρίσιμες΄, δηλαδή οι διαφορές μεταξύ τους παραμένουν μεγάλες, τότε είναι πιο ασφαλές να υποθέσουμε πως ένα μεγάλο spectral gap θα έχει σαν συνέπεια τα αποτελέσματα του αλγορίθμου HITS να συμπίπτουν με αυτά απο το μητρώο $e^{\mathcal{A}}$.

Αναφορές

[1] Michele Benzi, Ernesto Estrada, and Christine Klymko. Ranking hubs and authorities using matrix functions. *Linear Algebra and its Applications*, 438(5):2447–2474, 2013.

[2] Francesca Arrigo and Michele Benzi. Updating and downdating techniques for optimizing network communicability. *arXiv preprint arXiv:1410.5303*, 2014.

[3] David F. Gleich. Pagerank beyond the web. *Siam Review*, 57(3):321–363, 2015.

[4] Michele Benzi and Paola Boito. Matrix functions in network analysis. *Gamm-mitteilungen*, 43(3), 2020.