
Εθνικό Μετσόβιο Πολυτεχνείο

Σχολη Ηλεκτρολογων Μηχανικων
Και Μηχανικων Υπολογιστων

Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων


# Μεθοδοι καταταξης και μετρικες απηχησης επιστημονικων δημοσιευσεων

Διδακτορική Διατριβή

του


**Ηλία Κανέλλου**

Διπλωματούχου Ηλεκτρολόγου Μηχανικού & Μηχανικού Υπολογιστών
Εθνικού Μετσοβίου Πολυτεχνείου (2012)



Αθήνα, Ιανουάριος 2020

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων Και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

# Μέθοδοι κατάταξης και μετρικές απήχησης επιστημονικών δημοσιεύσεων

Διδακτορική Διατριβή

του

**Ηλία Κανέλλου**

Διπλωματούχου Ηλεκτρολόγου Μηχανικού & Μηχανικού Υπολογιστών
Εθνικού Μετσοβίου Πολυτεχνείου (2012)

**Συμβουλευτική Επιτροπή:** Ι. Βασιλείου
Τ. Σελλής
Θ. Δαλαμάγκας

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή την 15η Ιανουαρίου 2020.

| | | |
|---|---|---|
| Τ. Σελλής<br>Καθ. Swinburne University | Θ. Δαλαμάγκας<br>Ερευνητής Α΄<br>Ε. Κ. ΑΘΗΝΑ | Α. Γ. Σταφυλοπάτης<br>Καθ. ΕΜΠ |
| Χ. Παπαθεοδώρου<br>Καθ. Ιόνιο Παν/μιο | Σ. Σκιαδόπουλος<br>Καθ. Πανεπιστήμιο Πελοποννήσου | Δ. Γουνόπουλος<br>Καθ. ΕΚΠΑ |
| | Χ. Δουλκερίδης<br>Επ. Καθ. Πανεπιστήμιο Πειραιά | |

Αθήνα, Ιανουάριος 2020

...

**Ηλίας Κανέλλος**
Διδάκτωρ Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

# ΠΕΡΙΛΗΨΗ

Η διαρκής αύξηση του όγκου των επιστημονικών δημοσιεύσεων δημιουργεί προβλήματα στη διάκριση των σημαντικότερων από αυτές, επηρεάζοντας ταυτόχρονα ερευνητές, φοιτητές, υπεύθυνους ακαδημαϊκών προσλήψεων, αλλά και τις μηχανές αναζήτησης. Για το λόγο αυτό υπάρχει ανάγκη ανάπτυξης μηχανισμών κατάταξης (ή αλλιώς ιεράρχησης) των επιστημονικών δημοσιεύσεων. Παρ΄ ότι υπάρχει πλούσια βιβλιογραφία γύρω από την ανεξαρτήτως ερωτήματος κατάταξη (query independent ranking) επιστημονικών δημοσιεύσεων (γνωστή και ως στατική κατάταξη), στόχος της οποίας είναι η περιγραφή μεθόδων κατάταξης επιστημονικών δημοσιεύσεων, με βάση την απήχηση τους, δεν έχει πραγματοποιηθεί έως τώρα κάποια μεθοδική και συστηματική ανασκόπηση του αντικειμένου. Συγκεκριμένα, στη τρέχουσα βιβλιογραφία υφίστανται κενά στον ορισμό της απήχησης, ενώ δε γίνεται διάκριση μεταξύ μακροχρόνιας και βραχυχρόνιας επιστημονικής απήχησης. Επίσης, δεν εξετάζεται η σχέση της απήχησης των δημοσιεύσεων με άλλα χαρακτηριστικά των κειμένων, όπως η αναγνωσιμότητα. Επιπλέον, δεν έχει πραγματοποιηθεί καμία εκτενής πειραματική αξιολόγηση των επιμέρους μεθόδων που έχουν προταθεί στη βιβλιογραφία.

Αντικείμενο του διδακτορικού είναι η εξέταση της κατάταξης (ιεράρχησης) δημοσιεύσεων με βάση την απήχησή τους, των μεθόδων που έχουν προταθεί στη βιβλιογραφία, της εφαρμογής τους στα πλαίσια πραγματικών μηχανών αναζήτησης επιστημονικών δημοσιεύσεων και της συσχέτισης απήχησης - αναγνωσιμότητας δημοσιεύσεων. Συνοπτικά:

- Ορίζονται τυπικά η μακροχρόνια και βραχυχρόνια απήχηση και εξετάζονται και κατηγοριοποιούνται οι προσεγγίσεις που έχουν διατυπωθεί στη τρέχουσα βιβλιογραφία.

- Πραγματοποιείται μια εκτενής πειραματική αξιολόγηση για τη διερεύνηση των μηχανισμών που οδηγούν στη βέλτιστη αποτελεσματικότητα για την παραγωγή κατατάξεων των δημοσιεύσεων με βάση αυτά τα δύο είδη απήχησης.

- Καθώς τα αποτελέσματα της αξιολόγησης αποκαλύπτουν περιθώρια βελτίωσης στην κατάταξη με βάση τη βραχυχρόνια απήχηση, προτείνεται μια νέα μέθοδος κατάταξης, επηρεασμένη από πρόσφατες εξελίξεις της επιστήμης δικτύων (network science), ενσωματώνοντας τροποποιήσεις στη μέθοδο PageRank. Με εκτενή πειραματική αξιολόγηση αναδεικνύεται η αποτελεσματικότητα της νέας μεθόδου σε σχέση με τις άλλες τρέχουσες τεχνολογίες αιχμής.

- Παρουσιάζεται, επιπλέον, η ανάπτυξη εξειδικευμένων, αλλά και γενικών ακαδημαϊκών μηχανών αναζήτησης, οι οποίες κάνουν χρήση μεθόδων κατάταξης που έχουν εξεταστεί.

- Τέλος, εξετάζεται η σχέση της αναγνωσιμότητας των περιλήψεων επιστημονικών δημοσιεύσεων με την απήχησή τους.

**Λέξεις-κλειδιά:** Βιβλιομετρία, Κατάταξη Δημοσιεύσεων, Ανάκτηση Πληροφορίας, PageRank.

# ABSTRACT

The constantly increasing number of scientific publications affects researchers, students, academic hiring officials and search engines alike in discerning the high-impact works among them. Therefore, there is a need to develop methods to rank scientific papers. Despite a prolific literature on query-independent (or static) paper ranking algorithms, which aim to rank papers based on their impact, no systematic review of the field has been conducted. Past literature lacks in terms of defining impact, often failing to discern among short- term and long-term scientific impact. Further, no extensive experimental evaluation of the various proposed methods has been conducted.

This thesis examines impact-based paper ranking in terms of methods, search engine applications, and its relation to paper abstract readability. In short, the contributions of the thesis are as follows:

- Long-term and short-term impact are formally defined and the various ranking and evaluation approaches encountered so far in the literature are examined and classified.
- An extensive experimental evaluation is conducted to identify which proposed mechanisms perform best in ranking by short- and long-term impact.
- Motivated by the observed improvement margin in ranking based on short-term impact, a novel method is proposed building on recent advances of network science.
- The development of specialized and general academic search engines enhanced with short- and long-term impact-based ranking methods is presented.
- Finally, paper abstract readability and its relation to paper impact is examined.

**Keywords:** Bibliometrics, Paper Ranking, Information Retrieval, PageRank

# Περιεχόμενα

# Κατάλογος Σχημάτων

# Κατάλογος Πινάκων

# ΠΡΟΛΟΓΟΣ

*Η παρούσα διατριβή εκπληρώνει τις απαιτήσεις για την απόκτηση διπλώματος στο βαθμό του Διδάκτορα στη Σχολή Ηλεκτρολόγων και Μηχανικών Υπολογιστών στο Εθνικό Μετσόβειο Πολυτεχνείο (ΕΜΠ). Η εργασία που παρουσιάζεται περιγράφει μεθόδους κατάταξης και μετρικές απήχησης επιστημονικών δημοσιεύσεων και πραγματοποιήθηκε κατά τη διάρκεια των τελευταίων οκτώ χρόνων στο Εργαστήριο Βάσεων Γνώσεων και Δεδομένων του ΕΜΠ και στο Ινστιτούτο Πληροφοριακών Συστημάτων (ΙΠΣΥ) του Ερευνητικού Κέντρου «Αθηνά».*

*Είμαι ιδιαίτερα ευγνώμων στους Θανάση Βεργούλη και Θοδωρή Δαλαμάγκα για την εμπιστοσύνη που μου έδειξαν από την πρώτη στιγμή της συνεργασίας μας και την παρότρυνσή τους να συνεχίσω τις σπουδές μου. Στα πρόσωπά τους βρήκα εκτός από εξαιρετικούς μέντορες και δύο καλούς φίλους. Θέλω επίσης να ευχαριστήσω τον Δρ. Δημήτρη Σαχαρίδη για την πολύτιμη βοήθειά του, η οποία υπήρξε καθοριστική για την ολοκλήρωση της διατριβής. Τέλος, θέλω να ευχαριστήσω τους καθηγητές Τιμολέοντα Σελλή και Ιωάννη Βασιλείου για την εμπιστοσύνη που μου έδειξαν αναλαμβάνοντας την επίβλεψή μου. Θα ήθελα επίσης να ευχαριστήσω τους Ροδοθέα-Μυρσίνη Τσουπίδη, Κωσταντίνο Ζαγγανά, Βασιλική Βλαχοκυριάκου, Ανάργυρο Τζερεφό που συνεργάστηκαν μαζί μου ως προπτυχιακοί φοιτητές σε εργασίες σχετικές με το ερευνητικό μου αντικείμενο, καθώς και τον συνάδελφο υποψήφιο διδάκτορα Σεραφείμ Χατζόπουλο με τον οποίο είχα τη χαρά να συνεργάζομαι και να μοιράζομαι κοινά ερευνητικά ενδιαφέροντα.*

*Τέλος, θέλω να ευχαριστήσω θερμά όλους τους συναδέλφους στο Ε.Κ. «ΑΘΗΝΑ», η καθημερινότητα με τους οποίοθς δεν είναι απλά ευχάριστη, αλλά ενδιαφέρουσα και διασκεδαστική.*

*Ηλίας Κανέλλος*
*Αθήνα, Ιανουάριος 2020*

*Στο Γιάννη, την Εύα, τη Λένα και σε όσους ήταν κοντά μου αυτά τα χρόνια...*

# Κεφάλαιο 1

# Εισαγωγή

Οι επιστημονικές δημοσιεύσεις (Scientific Publications) αποτελούν, από την εποχή που ξεκίνησαν να εκδίδονται τα πρώτα επιστημονικά περιοδικά (Scientific Journals), το μηχανισμό μέσω του οποίου η επιστημονική γνώση που παράγεται από την έρευνα κατοχυρώνεται και διαδίδεται στην επιστημονική κοινότητα. Για να φτάσει μια επιστημονική δημοσίευση ως την έκδοση, περνάει από μια τυπική διαδικασία, κατά την οποία κρίνεται κατάλληλη για παρουσίαση σε κάποιο συνέδριο, ή περιοδικό, μέσω της αξιολόγησής της από ομότιμους (Peer Review). Στη διαδικασία αυτή άλλοι επιστήμονες κρίνουν κατά πόσο μια εργασία παρουσιάζει νέα γνώση, εντοπίζουν πιθανά σφάλματα στις υποθέσεις, στην πειραματική διαδικασία, ή στα συμπεράσματα της. Στη συνεχεία είτε αντιτίθενται στην έκδοσή της, είτε παραθέτουν προτάσεις βελτίωσης, ώστε μετά από διορθώσεις να δημοσιευτεί, είτε την προτείνουν για έκδοση χωρίς αλλαγές.

Η επιστημονική γνώση που παράγεται δεν αποτελεί ένα αυτοτελές και αποκομμένο από τον κόσμο προϊόν. Οι επιστήμονες στηρίζονται σε γνώση που έχει συσσωρευτεί προηγουμένως και έχει επίσης δημοσιευτεί σε επιστημονικές εργασίες. Οι συγγραφείς μιας έρευνας *αναφέρονται* (Cite) σε άλλες εργασίες, των οποίων κάνουν χρήση αποτελεσμάτων, συμπερασμάτων και μεθοδολογιών. Σήμερα η επιστημονική γνώση και η εξέλιξή της «καταγράφονται» σε ένα δίκτυο επιστημονικών *αναφορών* (References) μεταξύ των δημοσιεύσεων που εκδίδονται στα αναρίθμητα περιοδικά και συνέδρια ανά επιστημονικό κλάδο. Οι μηχανισμοί που παρατηρούνται καθώς αναπτύσσεται και εξελίσσεται η επιστημονική γνώση έχουν τέτοιο ενδιαφέρον που αποτελούν από μόνοι τους αντικείμενο ειδικών επιστημονικών κλάδων, της *βιβλιομετρίας* (Bibliometrics) και *επιστημομετρίας* (Scientometrics).

Επιπλέον, τις τελευταίες δεκαετίες ο ρυθμός με τον οποίο εκδίδονται επιστημονικές δημοσιεύσεις αυξάνεται διαρκώς και η τάση αυτή αναμένεται να συνεχιστεί [48, 9]. Μεγάλοι όγκοι δεδομένων γύρω από επιστημονικές δημοσιεύσεις (π.χ. κείμενα, αναφορές συμπληρωματικά υλικά) γίνονται πλέον διαθέσιμοι από διάφορους φορείς και πρωτοβουλίες «ανοιχτής επιστήμης» (Open Science), όπως για παράδειγμα τους BOAI[1], cOAlation S[2], και I4OC.[3] Επομένως, πέρα από τη βιβλιομετρία και την επιστημομετρία, τα δεδομένα αυτά αποκτούν ιδιαίτερο ενδιαφέρον και για την επιστήμη της πληροφορικής, ειδικά στην ανάκτηση πληροφορίας. Ήδη η αγορά ειδικών ακαδημαϊκών μηχανών αναζήτησης είναι μεγάλη: οι μηχανές Google Scholar, Microsoft Academic [68], Se-

[1] https://www.budapestopenaccessinitiative.org/

[2] https://www.scienceeurope.org/coalition-s/

[3] https://i4oc.org/

mantic Scholar[4], Pubmed[5], είναι μόνο ένα δείγμα των διαφόρων συστημάτων που καταγράφουν τους τεράστιους όγκους ακαδημαϊκών δημοσιεύσεων και λειτουργούν ως ακαδημαϊκές μηχανές αναζήτησης γενικού ή ειδικού σκοπού, κάθε μια από τις οποίες μπορεί να χρησιμοποιεί διαφορετικά υποσυστήματα και μηχανισμούς για να επιστρέφει αποτελέσματα σε χρήστες. Στα παραπάνω πλαίσια εντοπίζονται μια σειρά προβλημάτων και προκλήσεων, που αναλύονται στις επόμενες παραγράφους.

## 1.1 Προβλήματα και Προκλήσεις

### 1.1.1 Όγκος Αποτελεσμάτων Αναζήτησης

Παραδοσιακά, τα συστήματα αναζήτησης επιστημονικής βιβλιογραφίας χρησιμοποιούν κλασσικές τεχνικές της ανάκτησης πληροφορίας (Information Retrieval) για να κατατάξουν τα αποτελέσματα που εντοπίζουν ως σχετικά με μια αναζήτηση. Οι τεχνικές αυτές βασίζονται στην σχετικότητα των αποτελεσμάτων που επιστρέφονται με το εκάστοτε ερώτημα αναζήτησης (Query-dependent). Ωστόσο, ο αριθμός των δημοσιεύσεων που σχετίζονται με ένα αντικείμενο (κάποιον όρο αναζήτησης, ή λέξεις-κλειδιά), σήμερα, μπορεί να είναι πολύ μεγάλος. Αυτό μεταξύ των άλλων οφείλεται στο ότι οι επιστήμονες, υφίστανται διαρκώς πίεση να παραγάγουν δημοσιεύσεις, μια τάση γνωστή ως «δημοσιεύεις ή εξαφανίζεσαι» (Publish or Perish), η οποία έχει συσχετιστεί και με τη γενική πτώση της ποιότητας ενός μεγάλου μέρους των δημοσιεύσεων [65, 38]. Στην πράξη, κατά την αναζήτηση επιστημονικών δημοσιεύσεων, οι επιστήμονες δεν μπορούν να εξετάσουν εξαντλητικά τα αναρίθμητα σχετικά αποτελέσματα. Για το λόγο αυτό έχουν ανάγκη από συστήματα που τους επιτρέπουν εύκολα και γρήγορα να βρίσκουν τις «καλύτερες» δημοσιεύσεις που σχετίζονται με την αναζήτησή τους. Οι παραδοσιακές τεχνικές της ανάκτησης πληροφορίας δεν μπορούν να καλύψουν αυτό το κριτήριο, καθώς εστιάζουν μόνο στην σχετικότητα του περιεχομένου και όχι στη συνολική *απήχηση* των δημοσιεύσεων. Επομένως, οι παραδοσιακές τεχνικές πρέπει να συνδυαστούν και με άλλες μεθόδους, ανεξάρτητες του ερωτήματος (Query-independent), που στοχεύουν να κατατάξουν τις δημοσιεύσεις με βάση την *απήχηση* (Impact) τους.

### 1.1.2 Εξέλιξη Γράφων Αναφορών

Πολλές τεχνικές που έχουν προταθεί στη βιβλιογραφία για την κατάταξη επιστημονικών δημοσιεύσεων βασίζονται σε ανάλυση του γράφου που σχηματίζεται από τις αναφορές μεταξύ τους. Τυπικά χρησιμοποιούνται στοχαστικές διαδικασίες, όπως το PageRank [59] που μοντελοποιούν τη ροή της απήχησης που έχουν οι δημοσιεύσεις μεταξύ τους. Ωστόσο οι γράφοι αναφορών (Citation Graphs) εμπεριέχουν εγγενώς μεροληψία ενάντια σε πιο καινούριες δημοσιεύσεις [15, 37, 90, 51], διότι κάθε δημοσίευση μπορεί να αναφέρει μόνο άλλες που έχουν εκδοθεί νωρίτερα από την ίδια. Επομένως οι μέθοδοι κατάταξης που χρησιμοποιούνται πρέπει να εισαγάγουν μηχανισμούς άρσης αυτής της μεροληψίας, δεδομένου ότι και πολλές καινούριες ερευνητικές εργασίες μπορεί να έχουν μεγάλη επιστημονική απήχηση.

[4]https://www.semanticscholar.org/
[5]https://www.ncbi.nlm.nih.gov/pubmed/

### 1.1.3 Πληθώρα Μεθόδων Κατάταξης

Παρ' ότι υπάρχει πλούσια βιβλιογραφία σχετική με μεθόδους για κατάταξη επιστημονικών δημοσιεύσεων, ωστόσο το πεδίο δεν έχει ως τώρα εξεταστεί συστηματικά θεωρητικά, ή πειραματικά. Δεδομένου ότι πολλές από τις μεθόδους κατάταξης δημοσιεύσεων της βιβλιογραφίας προέρχονται από εργασίες που έγιναν σε διαφορετικούς επιστημονικούς κλάδους υπάρχουν μια σειρά από ζητήματα:

- Κάθε ερευνητική ομάδα μπορεί να αγνοεί δουλειά άλλων ομάδων κατά την αξιολόγηση των μεθόδων κατάταξης δημοσιεύσεων που προτείνει σε εργασίες της.
- Δεν υπάρχει κάποια ενιαία αποδεκτή μετρική για την αξιολόγηση της αποτελεσματικότητας μεθόδων κατάταξης δημοσιεύσεων, η οποία να αναγνωρίζεται από την ακαδημαϊκή κοινότητα [18].
- Οι σχετικές εργασίες χρησιμοποιούν διαφορετικά σύνολα δεδομένων για να κρίνουν την αποτελεσματικότητα μιας μεθόδου.
- Οι σχετικές εργασίες ακολουθούν διαφορετικές πειραματικές διαδικασίες για να εξετάσουν την αποτελεσματικότητα μιας μεθόδου.
- Συνολικά, δεν υπάρχει κάποιο ενιαίο πλαίσιο αξιολόγησής των μεθόδων κατάταξης. Όπως αναφέρθηκε πρόσφατα [2], αποτελεί ανοιχτό ζήτημα η ανάπτυξη πλαισίων αξιολόγησης σε κοινά σύνολα δεδομένων, πράγμα που θα επέτρεπε την ενιαία και αντικειμενική αξιολόγηση της απήχησης επιστημονικών δημοσιεύσεων.

Όλα τα παραπάνω δημιουργούν ένα κενό στην ανάπτυξη συστημάτων αναζήτησης δημοσιεύσεων, καθώς δεν είναι σαφές ποια μέθοδος θα έπρεπε να χρησιμοποιηθεί από ένα τέτοιο σύστημα και υπό ποιες συνθήκες. Η λύση του προβλήματος απαιτεί μια διεξοδική σύγκριση των μεθόδων της βιβλιογραφίας. Ως τώρα υπάρχουν μόνο λίγες και ανεπαρκείς σχετικές εργασίες: στο [67] συγκρίνονται μέθοδοι κατάταξης που όμως έχουν προταθεί για ιστοσελίδες του διαδικτύου. Η εργασία αυτή είναι σχετικά παρωχημένη και αγνοεί πολλές μεθόδους για κατάταξη επιστημονικών δημοσιεύσεων που έχουν προταθεί στη βιβλιογραφία. Επιπλέον, η εργασία αυτή χρησιμοποιεί ένα σύνολο δεδομένων από ένα μοναδικό επιστημονικό πεδίο και έτσι δε μπορεί να δώσει γενικευμένα συμπεράσματα. Μια δεύτερη εργασία [2] εξετάζει μόνο μια απλή κατηγοριοποίηση μεθόδων της βιβλιογραφίας, ενώ μια τρίτη [55] αγνοεί την ύπαρξη πολλών μεθόδων. Οι δύο τελευταίες επίσης δεν παρέχουν κάποια συγκριτική πειραματική αξιολόγηση.

### 1.1.4 Η Απήχηση δεν Είναι Μονοσήμαντη

Επιπλέον σημαντικό ζήτημα είναι το γεγονός ότι η απήχηση μιας δημοσίευσης «μπορεί να μετρηθεί ή να γίνει κατανοητή με πολλούς διαφορετικούς τρόπους» [8]. Για παράδειγμα, η παρούσα βιβλιογραφία αγνοεί το γεγονός ότι η απήχηση των δημοσιεύσεων μπορεί να μετρηθεί τόσο βραχυπρόθεσμα όσο και μακροπρόθεσμα. Για παράδειγμα ένας έμπειρος ερευνητής συνήθως αναζητάει *δημοφιλείς* δημοσιεύσεις, δηλαδή αυτές που αποτελούν το τρέχον σημείο αναφοράς της επιστημονικής κοινότητας στο πεδίο του. Από την άλλη μεριά, ένας νέος ερευνητής μπορεί να ενδιαφέρεται να βρει τις δημοσιεύσεις με τη μεγαλύτερη *επιρροή* στο επιστημονικό του πεδίο, οι οποίες το διαμόρφωσαν. Ακόμα θα μπορούσαν να εντοπιστούν διαφορετικά, επιπλέον χαρακτηριστικά που αντιστοιχούν και σε άλλα είδη απήχησης (π.χ. οι λεγόμενες εναλλακτικές μετρικές

- altmetrics μπορεί να θεωρηθούν ως κοινωνική απήχηση), η ποιοτική σημασία των οποίων και η μεταξύ τους σχέση θα πρέπει να εξερευνηθεί.

## 1.2 Συνεισφορά

Η συνεισφορά της διατριβής συνοψίζεται στα παρακάτω σημεία:

1. Ορίσαμε τυπικά τα προβλήματα κατάταξης δημοσιεύσεων με βάση δύο είδη απήχησης, την *επιρροή* και τη *δημοφιλία*, δυο έννοιες που στην τρέχουσα βιβλιογραφία δεν ορίζονται σαφώς και συχνά συγχέονται. Δώσαμε έτσι μια βάση για την ανάπτυξη υποβάθρων αληθείας, τα οποία μπορούν να χρησιμοποιηθούν στην αξιολόγηση μεθόδων κατάταξης.

2. Παρουσιάσαμε μια εκτενή βιβλιογραφική έρευνα για μεθόδους κατάταξης επιστημονικών δημοσιεύσεων που έχουν παρουσιαστεί σε πλήθος διαφορετικών επιστημονικών πεδίων. Οργανώνουμε τη βιβλιογραφία πραγματοποιώντας μια υψηλού επιπέδου κατηγοριοποίηση των μεθόδων, με βάση τα ποιοτικά χαρακτηριστικά τους, τις ιδέες τους και τους μηχανισμούς που χρησιμοποιούν.

3. Εξετάσαμε τις μεθοδολογίες που έχουν προταθεί στη βιβλιογραφία για την αξιολόγηση της αποτελεσματικότητας των μεθόδων κατάταξης, ώστε να ξεκαθαρίσουμε τους διαφορετικούς στόχους τους. Αναδείξαμε ταυτόχρονα την τρέχουσα έλλειψη μιας συγκριτικής αξιολόγησης με βάση ένα ενιαίο πλαίσιο.

4. Προτείναμε ένα συγκεκριμένο πλαίσιο αξιολόγησης με βάση το οποίο μπορούμε να ξεχωρίσουμε την αποτελεσματικότητά των διάφορων μεθόδων ως προς την παραγωγή κατατάξεων με βάση τη δημοφιλία και την επιρροή, δύο διακριτών χαρακτηριστικών που αντιστοιχούν στη βραχυχρόνια και τη μακροχρόνια επιστημονική απήχηση των δημοσιεύσεων.

5. Πραγματοποιήσαμε μια εκτεταμένη πειραματική αξιολόγηση μεθόδων κατάταξης με στόχο να μελετήσουμε: (α) πόσο διακριτές είναι οι έννοιες της δημοφιλίας και της επιρροής (β) ποιες μέθοδοι είναι αποτελεσματικότερες σε κάθε σενάριο κατάταξης και (γ) ποιες (επαναληπτικές) μέθοδοι συγκλίνουν γρηγορότερα. Αξιολογήσαμε ένα σύνολο μεθόδων επιλεγμένο έτσι ώστε να αντιπροσωπεύονται όλες οι διαφορετικές προσεγγίσεις που εντοπίσαμε στη σχετική βιβλιογραφία, σε 4 πραγματικά σύνολα δεδομένων, που ποικίλουν ως προς το μέγεθος και τον επιστημονικό κλάδο από τον οποίο προέρχονται, εξασφαλίζοντας έτσι ότι τα συμπεράσματά μας γενικεύονται.

6. Σχεδιάσαμε και αναπτύξαμε δύο πλήρως λειτουργικά συστήματα βιβλιογραφικής αναζήτησης: το BIP! Finder[6] και το mirPub v2.[7] Το BIP! Finder αποτελεί μια μηχανή αναζήτησης που δίνει στους χρήστες τη δυνατότητα να επιλέγουν το είδος απήχησης με βάση το οποίο θα ιεραρχούνται τα αποτελέσματα, καλύπτοντας έτσι διαφορετικές ανάγκες αναζήτησης σε σχέση με τις ακαδημαϊκές μηχανές αναζήτησης της αγοράς, οι οποίες στηρίζονται στην αρχή ότι ένα κριτήριο ιεράρχησης καλύπτει όλες τις ανάγκες. Το σύστημα καταγράφει 45 εκατομμύρια

[6]http://bip.imis.athena-innovation.gr

[7]http://mirpub.imis.athena-innovation.gr/projects/diana/index.php?r=mirpub

δημοσιεύσεις και παρέχει χρήσιμες οπτικοποιήσεις για τη σύγκριση των δημοσιεύσεων ως προς την απήχησή τους. Το mirPub v2 είναι μια μηχανή αναζήτησης βιβλιογραφίας σχετικής με microRNAs. Η ιδιαιτερότητα του συστήματος είναι ότι εφαρμόζει μηχανισμούς επέκτασης λέξεων-κλειδιών με βάση την ιστορική εξέλιξη των δεδομένων microRNA, ώστε να επιστρέφει την πληρέστερη δυνατή βιβλιογραφία, την οποία συνδυάζει και με βαθμολογίες κατάταξης. Οι χρήστες μπορούν να ιεραρχούν τα αποτελέσματα με βάση μια από τρεις μεθόδους που έχουν παραμετροποιηθεί για βέλτιστη αποτελεσματικότητα με βάση την επιρροή.

7. Σχεδιάσαμε και υλοποιήσαμε μια νέα μέθοδο κατάταξης βάσει της δημοφιλίας. Η μέθοδος μας ενσωματώνει ένα μηχανισμό «τρέχουσας προσοχής» στο μοντέλο του PageRank και πετυχαίνει πιο αποτελεσματικές κατατάξεις με βάση τη δημοφιλία σε σχέση με ανταγωνιστικές μεθόδους της βιβλιογραφίας.

8. Επεκτείνουμε τη θεώρηση ότι οι δημοσιεύσεις έχουν περισσότερα από ένα ενδιαφέροντα χαρακτηριστικά με βάση τα οποία μπορούν να εξεταστούν, πέρα από τη δημοφιλία και την επιρροή. Εξετάζουμε την αναγνωσιμότητα (Readability) τους και τη συσχέτισή της με την απήχηση από δύο οπτικές γωνίες: με βάση παραδοσιακές μετρικές αναγνωσιμότητας και με βάση τις (υποκειμενικές) κρίσεις ειδικών για την αναγνωσιμότητα.

9. Οι υλοποιήσεις των μεθόδων που αξιολογήσαμε πειραματικά αποτελούν ανοιχτή βιβλιοθήκη λογισμικού[8] (άδεια χρήσης GNU/GPL), ώστε να εξασφαλίζεται η αναπαραγωγιμότητα των πειραμάτων μας (στα πρότυπα της ανοιχτής επιστήμης), αλλά και να δοθεί κίνητρο για περαιτέρω έρευνα στο αντικείμενο από την επιστημονική κοινότητα. Οι υλοποιήσεις μας βασίζονται στο προγραμματιστικό μοντέλο Map-Reduce ώστε να εξασφαλίζουν την κλιμάκωση (Scalability) σε σενάρια μεγάλου όγκου δεδομένων.

## 1.3 Δομή της Διατριβής

Η υπόλοιπη έκθεση αναπτύσσεται ως εξής: στο Κεφάλαιο 2 δίνεται το υπόβαθρο στο οποίο βασίζεται η δουλειά και ορίζονται τυπικά τα προβλήματα κατάταξης με βάση τη μακροχρόνια και τη βραχυχρόνια απήχηση (επιρροή και δημοφιλία). Στο Κεφάλαιο 3 παρουσιάζεται μια αναλυτική επισκόπηση της παρούσας βιβλιογραφίας για μεθόδους ανεξάρτητης του ερωτήματος κατάταξης επιστημονικών δημοσιεύσεων, λεπτομερής κατηγοριοποίηση τους με βάση τις διάφορες προσεγγίσεις, και εξαντλητική εξέταση των μεθοδολογιών αξιολόγησης τους στην παρούσα βιβλιογραφία. Στο Κεφάλαιο 4 παρουσιάζεται ένα ενιαίο πλαίσιο αξιολόγησης με βάση το οποίο πραγματοποιείται η σύγκριση των τεχνολογιών αιχμής της παρούσας βιβλιογραφίας, ώστε να φανεί η συγκριτική αποτελεσματικότητα τους να παραγάγουν κατατάξεις με βάση την επιρροή και τη δημοφιλία. Στο Κεφάλαιο 5 παρουσιάζονται δύο πραγματικές μηχανές βιβλιογραφικής αναζήτησης που κάνουν χρήση μεθόδων κατάταξης, καθώς και μια νέα, δική μας μέθοδος που παράγει βελτιωμένες κατατάξεις με βάση τη δημοφιλία. Στο Κεφάλαιο 6 παρουσιάζουμε μια μελέτη της συσχέτισης της απήχησης των επιστημονικών δημοσιεύσεων με την αναγνωσιμότητα τους. Τέλος, στο Κεφάλαιο 7 συνοψίζουμε την παρουσίαση της εργασίας και καταγράφουμε τα επόμενα ερευνητικά βήματα.

[8] https://github.com/diwis

# Κεφάλαιο 2

# Βασικές Έννοιες Δικτύων Αναφορών και Απήχηση Δημοσιεύσεων

Στην ενότητα αυτή εισάγουμε κάποιες βασικές έννοιες σχετικές με τα Δίκτυα Αναφορών (Citation Networks), που χρησιμοποιούνται στα επόμενα κεφάλαια. Παρουσιάζουμε ορισμένα μέτρα κεντρικότητας (Centrality Metrics) και περιγράφουμε το πρόβλημα της κατάταξης δημοσιεύσεων. Επιπλέον ορίζουμε τυπικά την έννοια της Απήχησης (Impact), την οποία διακρίνουμε σε Μακροχρόνια (Long-term Impact) και Βραχυχρόνια (Short-term Impact) και, τέλος, παρουσιάζουμε τα προβλήματα κατάταξης που σχετίζονται με αυτές τις έννοιες.

## 2.1 Δίκτυα Αναφορών

*Δίκτυα Αναφορών*, ή *γράφους αναφορών* ονομάζουμε τους κατευθυνόμενους γράφους των οποίων οι κόμβοι αντιστοιχούν σε Επιστημονικές Δημοσιεύσεις (Scientific Publications) και των οποίων οι κατευθυνόμενες ακμές αντιστοιχούν σε αναφορές που γίνονται μεταξύ των επιστημονικών δημοσιεύσεων. Για κάθε κόμβο του δικτύου αναφορών, μια εξερχόμενη (εισερχόμενη) ακμή δηλώνει μια αναφορά που γίνεται προς (που λαμβάνεται από) μια άλλη δημοσίευση. Τα δίκτυα αναφορών είναι εξελισσόμενοι γράφοι, καθώς με την έκδοση νέων δημοσιεύσεων, εισάγονται νέοι κόμβοι στο γράφο και οι βαθμοί εισερχόμενων ακμών (In-degrees) των κόμβων ενδέχεται να αυξάνουν. Αντίθετα, οι βαθμοί εξερχόμενων ακμών (Out-degrees) μένουν πάντοτε αμετάβλητοι: κάθε κόμβος αντιστοιχεί σε μια επιστημονική δημοσίευση που έχει εκδοθεί, επομένως δεν μπορούν να τροποποιηθούν οι αναφορές που γίνονται από κείμενο της δημοσίευσης προς άλλες επιστημονικές δημοσιεύσεις.

Ένα δίκτυο αναφορών αποτελούμενο από $N$ δημοσιεύσεις περιγράφεται από τον $N \times N$ Πίνακα Γειτνίασης (Adjacency Matrix) $\boldsymbol{A}$, όπου $A_{i,j} = 1$ αν η δημοσίευση $j$ αναφέρει τη δημοσίευση $i$ (δηλαδή υπάρχει μια ακμή $j \to i$ στο δίκτυο αναφορών) και διαφορετικά $A_{i,j} = 0$. Για να αναφερθούμε στο χρόνο έκδοσης μιας επιστημονικής δημοσίευσης $i$, χρησιμοποιούμε το σύμβολο $t_i$. Με αυτό περιγράφεται το χρονικό σημείο εμφάνισης του κόμβου που συμβολίζει την εν λόγω δημοσίευση στο δίκτυο αναφορών. Η εισαγωγή του κάθε νέου κόμβου στο δίκτυο αναφορών συνοδεύεται φυσικά και από την εισαγωγή των κατευθυνόμενων ακμών από αυτόν τον κόμβο προς όλους τους άλλους που αντιστοιχούν σε δημοσιεύσεις που αναφέρει.

**Συμβάσεις Ορολογίας**. Στο εξής, για να αναφερθούμε στις επιστημονικές δημοσιεύσεις, θα χρησιμοποιούμε τον απλούστερο όρο «*δημοσιεύσεις*». Όταν σχολιάζουμε τη σχέση αναφοράς μεταξύ δύο δημοσιεύσεων $j \to i$ στο γράφο αναφορών, θα χρησιμοποιούμε τον όρο «*αναφέρουσα*» για να περιγράψουμε τη δημοσίευση $j$ που αναφέρει κάποια άλλη και τον όρο «*αναφερόμενη*» για να περιγράψουμε τη δημοσίευση $i$ που δέχεται την αναφορά. Στο πλαίσιο των δικτύων αναφοράς, όταν χρησιμοποιούμε την έννοια «*βαθμός*» θα αναφερόμαστε γενικά στον βαθμό εισερχόμενων ακμών (In-degree), δηλαδή τον αριθμό αναφορών που έχει λάβει μια δημοσίευση. Ισοδύναμα θα χρησιμοποιούμε την έννοια «*αριθμός αναφορών*». Όταν αναφερόμαστε στον βαθμό εξερχομένων ακμών αυτό θα δηλώνεται ρητά, είτε ως ο αριθμός αναφορών που *κάνει* μια δημοσίευση, είτε θα χρησιμοποιείται η έννοια του «*μεγέθους της λίστας αναφορών*» μιας δημοσίευσης $j$.

## 2.2 Μέτρα Κεντρικότητας

Στη θεωρία γράφων και στην ανάλυση δικτύων η έννοια της *κεντρικότητας* (Centrality) χρησιμοποιείται για να περιγράψει ορισμένους δείκτες σημαντικότητας των κόμβων που απαρτίζουν ένα γράφο. Η σημαντικότητα αυτή περιγράφεται από κάποια αριθμητική τιμή που αποδίδεται σε κάθε κόμβο, όπου μεγαλύτερες τιμές αντιστοιχούν σε σημαντικότερους κόμβους. Ακολούθως εξετάζουμε δύο *μέτρα κεντρικότητας* ιδιαίτερης σημασίας στα δίκτυα αναφορών: τον αριθμό αναφορών και το PageRank.

### 2.2.1 Αριθμός Αναφορών (Βαθμός Κόμβων)

Ο *βαθμός* ενός κόμβου $i$ αντιστοιχεί στον αριθμό αναφορών της δημοσίευσης την οποία αντιπροσωπεύει ο κόμβος. Ο βαθμός υπολογίζεται από τον πίνακα γειτνίασης ως $k_i = \sum_j A_{i,j}$. Συμβολίζουμε με $k_i^{out}$ τον βαθμό εξερχόμενων ακμών, δηλαδή το μέγεθος της λίστας αναφορών μιας δημοσίευσης $i$.

### 2.2.2 PageRank

Το PageRank [59, 10] προτάθηκε αρχικά στο πλαίσιο του διαδικτύου, με στόχο να αποδώσει κάποια ποσοτική τιμή σημαντικότητας σε κάθε ιστοσελίδα. Το PageRank αποτελεί ένα μέτρο κεντρικότητας που βασίζεται σε υπολογισμό ιδιοδιανύσματος πάνω σε έναν τροποποιημένο πίνακα γειτνίασης που περιγράφει το δίκτυο αναφορών, γνωστό στη βιβλιογραφία και ως πίνακα Google. Τα μέτρα κεντρικότητας που στηρίζονται σε υπολογισμό ιδιοδιανύσματος, γενικώς εκφράζουν την αρχή ότι «ένας κόμβος είναι σημαντικός εάν συνδέεται με άλλους σημαντικούς κόμβους».

Στο πλαίσιο των δικτύων αναφοράς το PageRank προσομοιώνει τη συμπεριφορά ενός «Τυχαίου Ερευνητή» (Random Researcher), μίας οντότητας που φέρεται ως εξής: ξεκινά τη δουλειά του διαβάζοντας κάποια δημοσίευση. Στη συνέχεια είτε επιλέγει να διαβάσει κάποια από τις αναφερόμενες δημοσιεύσεις από τη λίστα αναφορών αυτής που διαβάζει (διαλέγοντας οποιαδήποτε από αυτές με ίση πιθανότητα), είτε επιλέγει εντελώς τυχαία (με ίση πιθανότητα) να διαβάσει οποιαδήποτε δημοσίευση που βρίσκεται στο δίκτυο αναφορών. Η διαδικασία αυτή που ακολουθεί ο τυχαίος ερευνητής ονομάζεται *Τυχαία Περιήγηση* (Random Walk), ενώ η τυχαία επιλογή μιας δημοσίευσης που δεν βρέθηκε σε λίστα αναφορών ονομάζεται «τυχαίο άλμα», ή «τυχαία μετάβαση» (Random Jump). Αντίστοιχα στο διαδίκτυο η τυχαία περιήγηση αφορούσε ένα χρήστη που ξεκινά

από μια ιστοσελίδα και είτε ακολουθεί συνδέσμους προς άλλες σελίδες, είτε ανοίγει τυχαία μια οποιαδήποτε ιστοσελίδα. Το PageRank τελικά δίνει μια αριθμητική τιμή για κάθε δημοσίευση $i$, η οποία αντιστοιχεί στη βαθμολογία (Score) της δημοσίευσης. Η βαθμολογία PageRank κάθε δημοσίευσης $i$ δίνει την πιθανότητα να τη διαβάσει ένας τυχαίος ερευνητής κατά τη διάρκεια της τυχαίας περιήγησης (ή, ισοδύναμα, το ποσοστό χρόνου που αφιερώνει ο τυχαίος ερευνητής σε κάθε δημοσίευση $i$ κατά την τυχαία περιήγηση). Οι τιμές PageRank υπολογίζονται ως ακολούθως:

$$s_i = \alpha \sum_j P_{i,j} s_j + (1 - \alpha) v_i \tag{2.1}$$

Στην Εξίσωση 2.1, με $\boldsymbol{P}$ συμβολίζεται ο Πίνακας Μεταβάσεων (Transition Matrix) του δικτύου αναφορών, όπου $P_{i,j} = A_{i,j}/k_j^{out}$ (και $k_j^{out}$ ο αριθμός των αναφορών που κάνει η δημοσίευση $j$). Η τιμή $(1-\alpha)$ αντιστοιχεί στην Πιθανότητα Τυχαίου Άλματος (Random Jump Probability), η οποία ορίζει πόσο συχνά ο τυχαίος ερευνητής επιλέγει να διαβάσει στην τύχη μια οποιαδήποτε δημοσίευση του δικτύου αναφορών. Η τιμή $v_i$ ονομάζεται Πιθανότητα Επιλογής (Landing Probability) και ορίζει την πιθανότητα να επιλεγεί η δημοσίευση $i$ μετά από μια τέτοια τυχαία επιλογή. Συνήθως η πιθανότητα αυτή ορίζεται ίση για όλες τις δημοσιεύσεις και παίρνει την τιμή $v_i = 1/N$. Σημειώνεται ότι στην περίπτωση των «Εκκρεμών Κόμβων» (Dangling Nodes), εκείνων δηλαδή των κόμβων που δεν έχουν εξερχόμενες ακμές (δηλαδή δημοσιεύσεις οι οποίες δεν κάνουν αναφορές σε άλλες) ισχύει ότι $k_j^{out} = 0$. Συνεπώς, η τιμή του πίνακα μεταβάσεων δεν μπορεί να οριστεί. Στην περίπτωση αυτή, η συνήθης τεχνική είναι να τίθεται $P_{i,j} = A_{i,j}/N$ (ή κάποια άλλη πιθανότητα), όταν $k_j^{out} = 0$. Καθώς στην Εξίσωση 2.1 φαίνεται ότι η βαθμολογία κάθε κόμβου εξαρτάται από τις βαθμολογίες όλων των άλλων που τον αναφέρουν, οι υπολογισμοί για όλους τους κόμβους γίνονται επαναληπτικά μέχρι τη σύγκλιση. Μπορούμε στην Εξίσωση 2.1 να δώσουμε και μία άλλη διαισθητική περιγραφή του υπολογισμού, όπου κάθε δημοσίευση προωθεί ένα τμήμα της δικής της βαθμολογίας προς όλες τις αναφερόμενες από αυτήν δημοσιεύσεις. Σημειώνεται, τέλος, ότι ο τύπος του PageRank αποδίδει μια βαθμολογία σε κάθε δημοσίευση που εξαρτάται από όλα τα μονοπάτια που περνούν από αυτή.

Σε προηγούμενες εργασίες όπου γίνεται χρήση του PageRank σε δίκτυα αναφορών [15, 54], χρησιμοποιείται για το $\alpha$ η τιμή $\alpha = 0.5$. Η τιμή που συνηθίζεται, από την άλλη, για την κατάταξη ιστοσελίδων στο διαδίκτυο είναι $\alpha = 0.85$. Η παραδοχή που γίνεται είναι ότι κατά μέσο όρο ο τυχαίος ερευνητής ακολουθεί μια αναφορά από αυτές που αναγράφονται στη δημοσίευση που διαβάζει και στη συνέχεια επιλέγει τυχαία κάποια δημοσίευση από το δίκτυο αναφορών. Αντίθετα στο διαδίκτυο η τιμή $\alpha = 0.85$ βασίζεται στην παραδοχή ότι ένας τυχαίος περιηγητής ακολουθεί κατά μέσο όρο έξι φορές συνδέσμους προς άλλες ιστοσελίδες προτού ανοίξει τυχαία κάποια ιστοσελίδα.

## 2.3 Κατάταξη Επιστημονικών Δημοσιεύσεων

Στο επίκεντρο της παρούσας εργασίας βρίσκεται η έννοια της κατάταξης δημοσιεύσεων. Με τον όρο «κατάταξη» (Ranking) εννοούμε την παραγωγή μιας σειράς ιεράρχησης των δημοσιεύσεων που απαρτίζουν το γράφο αναφορών, με βάση κάποιο κριτήριο. Η σειρά ιεράρχησης προκύπτει μέσω της απόδοσης βαθμολογιών (Scores) στις δημοσιεύσεις, όπου μια δημοσίευση $i$ με βαθμολογία μεγαλύτερη από αυτή που αποδίδεται σε μια δημοσίευση $j$ κατατάσσεται σε πρότερη θέση κατάταξης από τη δεύτερη. Αν ορίσουμε

τη βαθμολογία της δημοσίευσης $i$ ως $s_i$ και τη θέση κατάταξης της δημοσίευσης $i$ ως $rank(i)$, τότε $s_i > s_j \implies rank(i) < rank(j)$. Αυτό υπονοεί ότι η δημοσίευση $i$ έχει μεγαλύτερη αξία από τη δημοσίευση $j$. Επομένως, εφόσον η δημοσίευση $i$ έχει μεγαλύτερη βαθμολογία κατάταξης, τότε σε μια *λίστα κατάταξης* (Ranked List), μία διάταξη δηλαδή των δημοσιεύσεων με την ιεραρχική τους σειρά, η δημοσίευση $i$ θα βρίσκεται ψηλότερα από τη δημοσίευση $j$.

Οι δημοσιεύσεις μπορεί να κατατάσσονται με βάση βαθμολογίες που υπολογίζονται με διάφορους τρόπους: για παράδειγμα, μπορεί να χρησιμοποιείται για την κατάταξη κάποιο από τα μέτρα κεντρικότητας που περιγράφηκαν στην Ενότητα 2.2. Στο Κεφάλαιο 3 θα εξετάσουμε εξαντλητικά τις διάφορες *μεθόδους κατάταξης* (Ranking Methods) που έχουν προταθεί στη βιβλιογραφία, με βάση τις διαφορετικές τους προσεγγίσεις και τους μηχανισμούς που υλοποιούν, ενώ στο Κεφάλαιο 4 εξετάζουμε την αποτελεσματικότητά τους στην κατάταξη δημοσιεύσεων με βάση την *απήχηση* τους, έννοια που ορίζουμε στη συνέχεια.

## 2.4 Απήχηση Επιστημονικών Δημοσιεύσεων

Για την κατάταξη των επιστημονικών δημοσιεύσεων με βάση την απήχηση μπορεί να γίνει χρήση μέτρων κεντρικότητας. Ωστόσο, δεν αρκεί ο απλός υπολογισμός της κεντρικότητας σε ένα γράφο αναφορών, διότι αυτό μπορεί να οδηγεί σε πολωμένα αποτελέσματα, π.χ. σε βάρος των νεότερων δημοσιεύσεων. Αυτό συμβαίνει εξαιτίας του συνδυασμού των παρακάτω:

- Όλα τα μέτρα κεντρικότητας βασίζονται στις ακμές του δικτύου (στις αναφορές που γίνονται μεταξύ των δημοσιεύσεων).
- Οι αναφορές μιας δημοσίευσης γίνονται πάντοτε προς άλλες δημοσιεύσεις που έχουν εκδοθεί νωρίτερα χρονικά σε σχέση με την αναφέρουσα δημοσίευση.
- Οι νεότερες δημοσιεύσεις δεν έχουν υπάρξει εντός του δικτύου για επαρκές χρονικό διάστημα ώστε να συσσωρεύουν αριθμό αναφορών που αντανακλά την πραγματική τους απήχηση.

Τα παραπάνω εγγενή χαρακτηριστικά των δικτύων αναφορών, έχουν ως βασικότερο επακόλουθο την εμφάνιση μιας καθυστέρησης από το χρόνο έκδοσης μιας δημοσίευσης μέχρι το χρόνο όπου αναφέρεται για πρώτη φορά από κάποια άλλη. Αυτό το φαινόμενο είναι γνωστό στη βιβλιογραφία ως «καθυστέρηση αναφοράς» (Citation Lag) [17, 6, 31, 70]. Επομένως η τρέχουσα απήχηση των δημοσιεύσεων στη πραγματικότητα φανερώνεται από την κεντρικότητά τους σε κάποια μέλλουσα χρονική στιγμή, όταν θα έχουν εκδοθεί νέες δημοσιεύσεις που τις αναφέρουν. Συνεπώς, η μέτρηση της απήχησης των δημοσιεύσεων με χρήση μέτρων κεντρικότητας θα πρέπει να λαμβάνει υπόψη την *εξέλιξη* των δικτύων αναφοράς.

Στις επόμενες υποενότητες ορίζουμε τυπικά δύο είδη απήχησης των δημοσιεύσεων, βασιζόμενοι στη χρήση μέτρων κεντρικότητας σε δίκτυα αναφορών. Τα είδη απήχησης που περιγράφουμε βασίζονται και τα δύο στη χρήση μελλοντικών καταστάσεων του δικτύου αναφορών. Σε όσα έπονται, ορίζουμε ως $A(t)$ το στιγμιότυπο του πίνακα γειτνίασης που περιγράφει ένα δίκτυο αναφορών, κατά τη χρονική στιγμή $t$, δηλαδή συμπεριλαμβάνοντας μόνο τις αναφορές μεταξύ δημοσιεύσεων που έχουν γίνει μέχρι εκείνη τη στιγμή. Επιπλέον ορίζουμε ως $t_c$ την τρέχουσα χρονική στιγμή.

### 2.4.1 Μακροχρόνια Απήχηση Δημοσιεύσεων

Το πρώτο είδος απήχησης αφορά το μακροχρόνιο αντίκτυπο μιας επιστημονικής δημοσίευσης και μπορούμε να πούμε ότι αντιστοιχεί στη *συνολική επιρροή* της δημοσίευσης στην επιστημονική κοινότητα. Ορίζουμε τη μακροχρόνια απήχηση (συνολική επιρροή) ως την κεντρικότητα μιας δημοσίευσης εάν θεωρήσουμε ως αναφορά κάποιο ιδεατό τελικό χρονικό σημείο [79] του πίνακα γειτνίασης του δικτύου αναφορών, ή αλλιώς την κεντρικότητα μιας δημοσίευσης στον πίνακα $\boldsymbol{A}(\infty)$.

### 2.4.2 Βραχυχρόνια Απήχηση Δημοσιεύσεων

Το δεύτερο είδος απήχησης αφορά το βραχυχρόνιο αντίκτυπο μιας επιστημονικής δημοσίευσης. Μπορούμε να πούμε ότι αυτό αντανακλά την τρέχουσα *δημοφιλία* της δημοσίευσης στην επιστημονική κοινότητα. Η δημοφιλία μπορεί να ποσοτικοποιηθεί μόνο χρήσει των αναφορών που θα πάρει κάθε δημοσίευση σε ένα άμεσο, κοντινό μέλλον. Με άλλα λόγια, μπορεί να μετρηθεί με βάση τις αναφορές που θα γίνουν σε μια δημοσίευση από άλλες που στην τρέχουσα χρονική στιγμή βασίζονται στην πρώτη, αλλά εκκρεμεί ακόμα η ολοκλήρωση και δημοσίευσή τους. Η απόσταση στην οποία πρέπει να τεθεί ο χρονικός ορίζοντας σε σχέση με την τρέχουσα χρονική στιγμή, για να μετρηθεί ορθά η δημοφιλία, εξαρτάται από τον τυπικό κύκλο παραγωγής της έρευνας σε κάθε επιστημονικό κλάδο (μελέτη-συγγραφή, κρίση, δημοσίευση). Αν θεωρήσουμε τη χρονική διάρκεια αυτή ως $T$, τότε η δημοφιλία μπορεί να δοθεί από κάποιο μέτρο κεντρικότητας στον πίνακα γειτνίασης $\boldsymbol{A}(t_c+T)-\boldsymbol{A}(t_c)$. Θεωρούμε στην περίπτωση αυτή ότι οι πίνακες $\boldsymbol{A}(t_c), \boldsymbol{A}(t_c+T)$ έχουν και οι δύο διάσταση $N \times N$, όπου $N$ ο συνολικός αριθμός των δημοσιεύσεων την χρονική στιγμή $t_c+T$. Περιέχουν ωστόσο μόνο τιμές που αντιστοιχούν στις αναφορές που γίνονται μέχρι τις χρονικές στιγμές $t_c, (t_c+T)$, αντίστοιχα. Συνεπώς ο πίνακας $\boldsymbol{A}(t_c+T)-\boldsymbol{A}(t_c)$ περιγράφει μόνο τις αναφορές που πραγματοποιήθηκαν στο χρονικό διάστημα $[t_c, t_c+T]$.

Εάν επιλέξουμε ως μέτρο κεντρικότητας το βαθμό των κόμβων, τότε ως δημοφιλία ορίζουμε τον αριθμό άμεσων αναφορών μιας δημοσίευσης από άλλες σε ένα χρονικό διάστημα. Αν από την άλλη χρησιμοποιήσουμε το PageRank, τότε βασιζόμαστε και σε όλα τα μονοπάτια αναφορών που περνάνε από τον κάθε κόμβο με βάση τις αναφορές που γίνονται σε ένα δοσμένο χρονικό διάστημα.

### 2.4.3 Κατάταξη Βάσει Απήχησης

Για τη συγκριτική μελέτη της αποτελεσματικότητας μεθόδων κατάταξης επιστημονικών δημοσιεύσεων βάσει της απήχησης (θα εξεταστεί στο Κεφάλαιο 4), ορίζουμε δύο προβλήματα κατάταξης: κατάταξη βάσει βραχυχρόνιας και κατάταξη βάσει μακροχρόνιας απήχησης. Έχοντας ορίσει στις Ενότητες 2.4.1 και 2.4.2 τη μακροχρόνια και βραχυχρόνια απήχηση (ή τη συνολική επιρροή και δημοφιλία, αντίστοιχα), τα προβλήματα αυτά μπορούν να οριστούν ως ακολούθως:

**Πρόβλημα 1. - Κατάταξη βάσει μακροχρόνιας απήχησης (συνολικής επιρροής)**: *Δεδομένης της τρέχουσας κατάστασης ενός δικτύου αναφορών (στο χρόνο $t_c$), να παραχθεί μια κατάταξη δημοσιεύσεων που να προσεγγίζει την κατάταξή τους με βάση την συνολική τους επίδραση, δηλαδή την αναμενόμενη κεντρικότητά τους στην κατάσταση $\boldsymbol{A}(\infty)$.*

**Πρόβλημα 2. - Κατάταξη βάσει βραχυχρόνιας απήχησης (δημοφιλίας)**: *Δεδομένης της κατάστασης του δικτύου αναφορών στην τρέχουσα χρονική στιγμή $t_c$, να παραχθεί μια κατάταξη που να προσεγγίζει την κατάταξή τους βάσει της δημοφιλίας, δηλαδή την αναμενόμενη κεντρικότητά τους στον πίνακα $\boldsymbol{A}(t_c+T)-\boldsymbol{A}(t_c)$, όπου $T$ είναι μια παράμετρος του προβλήματος.*

# Κεφάλαιο 3

# Επισκόπηση Τεχνολογιών Αιχμής Μεθόδων Κατάταξης Δημοσιεύσεων

Σε αυτή την ενότητα παρουσιάζουμε μια επισκόπηση των τεχνολογιών αιχμής μεθόδων κατάταξης (Ranking Methods) επιστημονικών δημοσιεύσεων, που έχουν προταθεί στη βιβλιογραφία και κάνουν χρήση δικτύων αναφορών. Εξετάζουμε τις πιο σημαντικές ιδέες που προτείνονται στη βιβλιογραφία και τον τρόπο με τον οποίο εφαρμόζονται στις επιμέρους μεθόδους κατάταξης. Η κατηγοριοποίηση γίνεται με βάση δύο άξονες: πρώτον, τη χρήση, ή μη, χρονικών παραγόντων. Δεύτερον, τη χρήση των ειδών συμπληρωματικής πληροφορίας πέρα από τους κόμβους και τις ακμές των δικτύων αναφορών. Με βάση τον πρώτο άξονα διακρίνουμε μεθόδους που δεν χρησιμοποιούν χρονικούς παράγοντες, οι οποίες αποτελούν κυρίως απλές παραλλαγές του PageRank (PR), και μεθόδους που εισάγουν χρονικές παραμέτρους σε πίνακες γειτνίασης/μεταβάσεων, ή στις πιθανότητες επιλογής του PageRank. Ο δεύτερος άξονας που εξετάζεται, σχετίζεται με τη χρήση άλλων μεταδεδομένων των δημοσιεύσεων (συγγραφείς, περιοδικά, συνέδρια), ή την ανάλυση πολλαπλών δικτύων (συγγραφείς-δημοσιεύσεις, δημοσιεύσεις-περιοδικά). Επιπλέον στο κεφάλαιο αυτό εξετάζουμε συνδυαστικές μεθόδους, καθώς και μεθόδους που δεν εμπίπτουν στις παραπάνω κατηγοριοποιήσεις. Μια σύνοψη της συνολικής κατηγοριοποίησης των μεθόδων που εξετάζουμε παρουσιάζεται στον Πίνακα 3.1

## 3.1 Κατηγοριοποίηση Μεθόδων Κατάταξης

### 3.1.1 Βασικές Παραλλαγές PageRank

Αναφερόμαστε εδώ σε παραλλαγές που εφαρμόζονται στον πίνακα μεταβάσεων $\boldsymbol{P}$ του PageRank, οι οποίες δεν χρησιμοποιούν μεταδεδομένα, ή χρονικές παραμέτρους. Στην κατηγορία αυτή ανήκει το Non-Linear PageRank [87], μια μέθοδος βασισμένη στο PageRank, όπου εισάγονται όμως μη γραμμικοί υπολογισμοί σε μια παραλλαγή της Εξίσωσης 2.1. Συγκεκριμένα η μέθοδος τροποποιεί το πρώτο μέλος της δεξιά πλευράς της εξίσωσης, αθροίζοντας για κάθε δημοσίευση τις τιμές Non-Linear PageRank των αναφερουσών δημοσιεύσεων, υψωμένες σε κάποια δύναμη $\theta \in (0, 1)$, και υπολογίζοντας τελικά την $\theta$-ρίζα του αποτελέσματος. Ο σκοπός είναι κατά τον στον υπολογισμό των τελικών τιμών να έχουν περισσότερη επίδραση οι «σημαντικές» αναφέρουσες δη-

μοσιεύσεις και μικρότερη επίδραση οι λιγότερο σημαντικές αναφέρουσες δημοσιεύσεις.

Η μέθοδος SPRank [95] ενσωματώνει ένα «μηχανισμό προτίμησης» στη μοντελοποίηση του τυχαίου ερευνητή, τροποποιώντας τον πίνακα μεταβάσεων του PageRank. Ο μηχανισμός προτίμησης βασίζεται στην ομοιότητα μεταξύ της αναφέρουσας και της αναφερόμενης δημοσίευσης. Συγκεκριμένα η ομοιότητα υπολογίζεται με βάση την τομή των αναφορών που κάνουν η αναφέρουσα και η αναφερόμενη δημοσίευση προς τρίτες δημοσιεύσεις. Έτσι ο τυχαίος ερευνητής μετατρέπεται σε *σκόπιμο ερευνητή* (Focused Researcher), ο οποίος προτιμά, όταν διαλέγει μια δημοσίευση από τη λίστα αναφορών εκείνης που εξετάζει, να επιλέγει τις πιο όμοιες με αυτή που διαβάζει.

Μια άλλη προσέγγιση δίνεται από τη μέθοδο SCEAS [67], στην οποία η κάθε τιμή του πίνακα μεταβάσεων του PageRank πολλαπλασιάζεται με μια ποσότητα $q \in (0, 1)$. Το αποτέλεσμα αυτής της τροποποίησης είναι να μειώνεται δραστικά η συμβολή μεγαλύτερων μονοπατιών αναφορών που οδηγούν σε κάθε δημοσίευση, κατά τον υπολογισμό της βαθμολογίας της.

Η μέθοδος PrestigeRank [72] δεν τροποποιεί το PageRank καθεαυτό, αλλά το δίκτυο πάνω στο οποίο υπολογίζεται. Συγκεκριμένα, η μέθοδος προσθέτει έναν εικονικό κόμβο στο δίκτυο, ο οποίος αντιπροσωπεύει όλες τις δημοσιεύσεις που μπορεί να υπάρχουν στο πραγματικό δίκτυο αναφορών, αλλά οι οποίες δεν περιλαμβάνονται στο διαθέσιμο σύνολο δεδομένων (Dataset) που το αντιπροσωπεύει και το οποίο χρησιμοποιείται για τον υπολογισμό. Σκοπός αυτής της τροποποίησης, επομένως, είναι να παράξει μια «δίκαιη» κατάταξη σε περιπτώσεις ελλιπών δεδομένων. Για να κατανοηθεί η «αδικία» που η μέθοδος επιχειρεί να αντιμετωπίσει ας εξετάσουμε την ακόλουθη περίπτωση: όταν περιλαμβάνεται στο σύνολο δεδομένων μόνο ένα τμήμα της λίστας αναφορών μιας δημοσίευσης, τότε προωθείται μια μεγαλύτερη βαθμολογία από τη πραγματική από την αναφέρουσα δημοσίευση στις αναφερόμενες. Αυτό συμβαίνει καθώς ο πίνακας μεταβάσεων $\boldsymbol{P}$ που μοντελοποιεί το σύνολο δεδομένων χρησιμοποιεί μια τιμή $k_j^{out}$ μικρότερη από την πραγματική, με αποτέλεσμα η τιμή $1/k_j^{out}$ της Εξίσωσης 2.1 να αυξάνει. Αυτό επηρεάζει και τα τελικά αθροίσματα που υπολογίζονται για κάθε δημοσίευση.

Τέλος, η μέθοδος Focused PageRank [46] τροποποιεί τον πίνακα μεταβάσεων ως εξής: στη θέση του αντίστροφου του αριθμού των αναφερόμενων δημοσιεύσεων $1/k_j^{out}$ χρησιμοποιεί για κάθε αναφερόμενη δημοσίευση το ποσοστό των αναφορών που γίνονται σε αυτή μέσα στο κείμενο της αναφέρουσας δημοσίευσης. Ο σκοπός της συγκεκριμένης τροποποίησης είναι να διαφοροποιήσει τις αναφερόμενες δημοσιεύσεις με βάση το βαθμό στον οποίο έχουν επηρεάσει την αναφέρουσα.

### 3.1.2 Μέθοδοι Κατάταξης με Χρήση Χρονικών Παραγόντων

Σ' αυτή την υποενότητα παρουσιάζουμε τους διάφορους τρόπους χρήσης χρονικών παραγόντων που έχουν χρησιμοποιηθεί από τις μεθόδους κατάταξης δημοσιεύσεων στη βιβλιογραφία. Οι προσεγγίσεις που ακολουθούνται αφορούν κατά βάση την τροποποίηση των πινάκων γειτνίασης, ή μεταβάσεων του δικτύου αναφορών, ή και την τροποποίηση των πιθανοτήτων επιλογής, όταν πρόκειται για παραλλαγές του PageRank.

#### 3.1.2.1 Χρονικοί Παράγοντες στον Πίνακα Γειτνίασης/Μετάβασης.

Οι πίνακες γειτνίασης, ή μετάβασης του δικτύου αναφορών μπορούν να τροποποιηθούν με χρήση τριών διαφορετικών χρονικών ποσοτήτων. Οι ποσότητες αυτές συμβολίζονται εδώ με $\tau_{ij}$ όταν αφορούν τη σχέση αναφοράς μεταξύ δύο δημοσιεύσεων $j \to i$. Διακρίνουμε τρεις περιπτώσεις ορισμού του $\tau_{ij}$.

- *Ηλικία Αναφοράς*: αναφέρεται στην ηλικία της αναφέρουσας δημοσίευσης $j$ και ορίζεται ως $\tau_{ij} = t - t_j$.
- *Χρονικό Διάστημα Αναφοράς*: αναφέρεται στο χρόνο που έχει μεσολαβήσει από την έκδοση της αναφερόμενης δημοσίευσης $i$, μέχρι την αναφορά της από την αναφέρουσα δημοσίευση $j$ και ορίζεται ως $\tau_{ij} = t_j - t_i$.
- *Ηλικία Αναφερόμενης Δημοσίευσης*: αναφέρεται στην ηλικία της αναφερόμενης δημοσίευσης $i$ και ορίζεται ως $\tau_{ij} = t - t_i$.

Ο επικρατέστερος τρόπος με τον οποίο εισάγονται οι παραπάνω ποσότητες σε μεθόδους κατάταξης, είναι με την τροποποίηση του πίνακα γειτνίασης, έτσι ώστε κάθε μη μηδενικό στοιχείο του να αντικαθίσταται από μια εκθετική φθίνουσα συνάρτηση του $\tau_{ij}$:

$$A'_{i,j} = \kappa e^{-\gamma \tau_{ij}} A_{i,j},$$

όπου $\gamma > 0$ είναι ο *ρυθμός φθοράς* (Decay Rate) και το $\kappa$ κωδικοποιεί επιπλέον παράγοντες ή/και όρους κανονικοποίησης.

Όταν ως $\tau_{ij}$ χρησιμοποιείται η ηλικία αναφοράς, ουσιαστικά δίνεται μεγαλύτερο βάρος σε αναφορές που έχουν γίνει προς δημοσιεύσεις πρόσφατα. Όταν χρησιμοποιείται ως $\tau_{ij}$ το χρονικό διάστημα αναφοράς, τότε δίνεται μεγαλύτερο βάρος σε δημοσιεύσεις που έλαβαν αναφορές πολύ σύντομα μετά την έκδοσή τους. Τέλος, όταν χρησιμοποιείται ως $\tau_{ij}$ η ηλικία της αναφερόμενης δημοσίευσης, δίνεται μεγαλύτερο βάρος σε αναφορές προς δημοσιεύσεις που έχουν εκδοθεί πρόσφατα. Παρ' ότι είναι εφικτή η συνδυαστική χρήση των χρονικών ποσοτήτων που περιγράψαμε, στη παρούσα βιβλιογραφική επισκόπηση δεν βρέθηκε να υπάρχει προσέγγιση που να συνδυάζει και τις τρεις.

Η επίδραση ενός χρονικά ενήμερου (Time-Aware) πίνακα γειτνίασης στο βαθμό των κόμβων είναι άμεση: το άθροισμα $\sum_j A'_{i,j}$ υποδηλώνει ένα *σταθμισμένο αριθμό αναφορών*. Η προσέγγιση αυτή ακολουθείται από τις μεθόδους Weighted Citation [84] και MR-Rank [93], οι οποίες κάνουν χρήση του χρονικού διαστήματος αναφοράς. Επίσης, η μέθοδος Retained Adjacency Matrix [28] ακολουθεί την ίδια προσέγγιση χρησιμοποιώντας την ηλικία αναφοράς.

Στην περίπτωση της τροποποίησης του πίνακα μεταβάσεων σε παραλλαγές του PageRank, η αξία μιας αναφοράς δεν εξαρτάται μόνο από τη δημοσίευση που την κάνει, αλλά και από τη σχετική μ' αυτή χρονική ποσότητα $\tau_{ij}$ που χρησιμοποιείται. Η ιδέα αυτή υιοθετείται από τις μεθόδους κατάταξης Timed PageRank [89, 90] και NewRank [18], με χρήση της ηλικίας αναφερόμενης δημοσίευσης. Οι μέθοδοι αυτοί υπολογίζουν την βαθμολογία μιας δημοσίευσης ως:

$$s_i = \alpha \sum_j \kappa e^{-\gamma \tau_{ij}} P_{i,j} s_j + (1 - \alpha) v_i \tag{3.1}$$

Μια άλλη μέθοδος που χρησιμοποιεί χρονικές παραμέτρους και συγκεκριμένα την ηλικία αναφοράς, είναι η Effective Contagion Matrix [28]. Η μέθοδος αυτή ωστόσο δεν

τροποποιεί τον πίνακα μεταβάσεων του PageRank, αλλά τον πίνακα γειτνίασης. Στην πραγματικότητα η μέθοδος αποτελεί τροποποίηση ενός άλλου μέτρου κεντρικότητας, της κεντρικότητας Katz [51]. Το μέτρο αυτό, όπως και το PageRank, στηρίζεται σε υπολογισμό ιδιοδιανύσματος και λαμβάνει υπόψη όλα τα μονοπάτια που οδηγούν προς κάθε κόμβο.

Εκτός από τις προαναφερθείσες χρονικές ποσότητες έχουν προταθεί σε περιορισμένη βιβλιογραφία και κάποιες άλλες. Για παράδειγμα στο [16] χρησιμοποιούνται βάρη βασισμένα στο λόγο του συνολικού αριθμού αναφορών μιας δημοσίευσης προς την ηλικία της (κάτι που μπορεί να ερμηνευθεί ως ο μέσος αριθμός αναφορών που λαμβάνει κάθε έτος). Επιπλέον, στο [53] και την επέκτασή του που ονομάζεται SARank [55] χρησιμοποιούνται βάρη που προσδιορίζονται από εκθετικά φθίνουσες συναρτήσεις της ηλικίας μιας δημοσίευσης, αλλά μόνο στην περίπτωση που η δημοσίευση έχει φτάσει σε κάποια κορύφωση όσον αφορά τις αναφορές που λαμβάνει, δηλαδή έχει ξεπεράσει το έτος στο οποίο έλαβε το μέγιστο αριθμό αναφορών.

#### 3.1.2.2 Χρονικοί Παράγοντες στις Πιθανότητες Επιλογής

Στην κλασσική μορφή του PageRank και σε πολλές από τις παραλλαγές του, γίνεται η θεώρηση πως όλες οι δημοσιεύσεις έχουν την ίδια πιθανότητα επιλογής κατά τη πραγματοποίηση ενός τυχαίου άλματος. Ωστόσο κάποιες μέθοδοι αποδίδουν άνισες πιθανότητες επιλογής μεταξύ των δημοσιεύσεων. Συμβολίζουμε με $v_i$ την πιθανότητα επιλογής που ανατίθεται στη δημοσίευση $i$. Οι μέθοδοι που έχουν προταθεί ως τώρα στη βιβλιογραφία χρησιμοποιούν πιθανότητες επιλογής εκθετικά φθίνουσες με την ηλικία της κάθε δημοσίευσης, στη γενική μορφή:

$$v_i = \kappa e^{-\gamma(t-t_i)}$$

Η μορφή αυτή υπονοεί ότι πιο πρόσφατες δημοσιεύσεις έχουν μεγαλύτερη αξία από παλαιότερες. Σημειώνεται ότι σε αντίθεση με την προηγούμενη ενότητα, όπου περιγράψαμε χρονικές ποσότητες που αφορούν τις *ακμές* του δικτύου αναφοράς, εδώ περιγράφονται χρονικές ποσότητες που αφορούν την ηλικία μιας δημοσίευσης, δηλαδή τους *κόμβους* του δικτύου αναφορών.

Στη παρούσα βιβλιογραφία μπορούμε να διακρίνουμε δύο τρόπους παραλλαγής των πιθανοτήτων επιλογής. Ο πρώτος αφορά τις πιθανότητες τυχαίας μετάβασης του PageRank, όπως γίνεται στην περίπτωση του θεματικά προσδιορισμένου (Topic Sensitive) [34], ή προσωποποιημένου (Personalized) [39] PageRank. Μέθοδοι που χρησιμοποιούν χρονικές παραμέτρους κατα αυτόν τον τρόπο είναι οι CiteRank[78], FutureRank [66], YetRank [37], και NewRank [18]. Οι μέθοδοι αυτοί υπολογίζουν την βαθμολογία κάθε δημοσίευσης με ένα τύπο της μορφής:

$$s_i = \alpha \sum_j P_{i,j} s_j + (1-\alpha)\kappa e^{-\gamma(t-t_i)}.$$

Στη πραγματικότητα η μέθοδος NewRank χρησιμοποιεί επίσης ένα χρονικά ενήμερο πίνακα μεταβάσεων, ενώ οι υπολογισμοί της μεθόδου FutureRank εμπλέκουν επιπλέον ένα δεύτερο δίκτυο που περιγράφεται από ένα πίνακα συσχέτισης συγγραφέων με δημοσιεύσεις, όπως παρουσιάζεται στην Ενότητα 3.1.4.

Ο δεύτερος τρόπος παραλλαγής των πιθανοτήτων επιλογής αφορά στους εκκρεμείς κόμβους. Η τυπική προσέγγιση είναι να ορίζονται τεχνητά ακμές από τους εκκρεμείς κόμβους προς όλους τους άλλους κόμβους του δικτύου. Αντίθετα όμως, η μέθοδος

YetRank [37] χρησιμοποιεί εκθετικά φθίνουσες πιθανότητες μετάβασης από τους εκκρεμείς κόμβους, σύμφωνα με την εξίσωση:

$$P_{i,j} = \kappa e^{-\gamma(t-t_i)}, \text{ για κάθε } i,j : k_j^{out} = 0.$$

### 3.1.3 Μέθοδοι Κατάταξης με Χρήση Μεταδεδομένων

Μια άλλη κατηγορία μεθόδων κατάταξης δημοσιεύσεων χρησιμοποιεί τα μεταδεδομένα των δημοσιεύσεων, όπως τους συγγραφείς, το περιοδικό, ή συνέδριο στο οποίο δημοσιεύτηκαν. Οι βαθμολογίες που υπολογίζονται από τέτοιες μεθόδους μπορεί να βασίζονται σε απλά στατιστικά (π.χ. στους μέσους όρους των βαθμολογιών των δημοσιεύσεων ενός συγγραφέα), ή σε άλλα γνωστά μέτρα που έχουν προταθεί στη βιβλιογραφία όπως για παράδειγμα τον Impact Factor [26] ή τον Eigenfactor [5] που χρησιμοποιούνται για κατάταξη επιστημονικών περιοδικών. Η πλειοψηφία των μεθόδων που κάνουν χρήση μεταδεδομένων, ενσωματώνουν πληροφορία σχετικά με αυτά στον τύπο του PageRank τροποποιώντας τον πίνακα μεταβάσεων, ή τις πιθανότητες επιλογής. Μια άλλη δυνατότητα είναι η τροποποίηση του πίνακα γειτνίασης.

Για παράδειγμα, η μέθοδος Weighted Citation τροποποιεί τον πίνακα γειτνίασης $\boldsymbol{A}$ χρησιμοποιώντας βάρη που εξαρτώνται από το περιοδικό στο οποίο εκδόθηκαν οι αναφέρουσες δημοσιεύσεις. Επομένως η μέθοδος δίνει μεγαλύτερη αξία σε δημοσιεύσεις που αναφέρονται από άλλες, οι οποίες εκδόθηκαν σε περιοδικά, ή συνέδρια υψηλού κύρους.

Η μέθοδος YetRank [37] τροποποιεί τον πίνακα μεταβάσεων $\boldsymbol{P}$ του PageRank, καθώς και τις πιθανότητες επιλογής $\boldsymbol{v}$. Πιο συγκεκριμένα, η μέθοδος κάνει χρήση του Impact Factor στον καθορισμό της πιθανότητας επιλογής μιας δημοσίευσης κατά την εκκίνηση μιας τυχαίας περιήγησης, ή κατά την επιλογή μιας επόμενης δημοσίευσης μετά την επίσκεψη σε έναν εκκρεμή κόμβο. Κατ΄ αυτόν τον τρόπο, ή μέθοδος προσομοιώνει ερευνητές που προτιμούν να διαβάζουν δημοσιεύσεις που παρουσιάστηκαν σε περιοδικά, ή συνέδρια μεγάλου κύρους.

Η μέθοδος NTUWeightedPR [16] επίσης τροποποιεί τον πίνακα μεταβάσεων $\boldsymbol{P}$ και τις πιθανότητες επιλογής $\boldsymbol{v}$. Χρησιμοποιεί βάρη που εξαρτώνται από τους συγγραφείς, το περιοδικό και το ρυθμό λήψης αναφορών (Citation Rate). Με τον τρόπο αυτό η μέθοδος προσομοιώνει έναν «σκόπιμο» ερευνητή που προτιμάει να ακολουθεί αναφορές προς δημοσιεύσεις που έχουν συγγραφεί από γνωστούς συγγραφείς, ή έχουν δημοσιευτεί σε υψηλού κύρους περιοδικά, ή λαμβάνουν πολλές αναφορές κάθε έτος. Οι ίδιες προτιμήσεις στη συμπεριφορά του ερευνητή που προσομοιώνεται ισχύουν και στην περίπτωση τυχαίας μετάβασης προς οποιαδήποτε δημοσίευση.

Μια εναλλακτική προσέγγιση στα παραπάνω προτείνεται στη μέθοδο Timed PageRank [89, 90], η οποία υπολογίζει ξεχωριστά τις βαθμολογίες των πρόσφατων δημοσιεύσεων, για τις οποίες υπάρχει μόνο περιορισμένη πληροφορία σχετικά με τις αναφορές που έχουν λάβει. Συγκεκριμένα, για τις δημοσιεύσεις αυτές χρησιμοποιούνται μόνο τα μεταδεδομένα στον υπολογισμό βαθμολογιών, ενώ χρησιμοποιείται μια εκδοχή του PageRank με χρονικές παραμέτρους για τις υπόλοιπες. Για τις πρόσφατες δημοσιεύσεις χρησιμοποιούνται απλά στατιστικά, όπως οι μέσοι όροι των υπολοίπων δημοσιεύσεων του ίδιου συγγραφέα, ή βαθμολογίες που βασίζονται στο μέσο όρο των βαθμολογιών άλλων δημοσιεύσεων που εκδόθηκαν στο ίδιο περιοδικό, ή συνέδριο.

### 3.1.4 Μέθοδοι Κατάταξης με Χρήση Πολλαπλών Δικτύων

Μια άλλη κατηγορία μεθόδων κατάταξης δημοσιεύσεων κάνει χρήση πολλαπλών δικτύων πάνω στα οποία εφαρμόζονται επαναληπτικοί υπολογισμοί. Παραδείγματα τέτοιων δικτύων είναι τα δίκτυα συγγραφέων-δημοσιεύσεων, περιοδικών-δημοσιεύσεων κλπ. Τα δίκτυα αυτά μπορεί να χρησιμοποιούνται συμπληρωματικά με το κλασσικό δίκτυο αναφορών. Διακρίνουμε δύο υποκατηγορίες: η πρώτη υποκατηγορία χρησιμοποιεί την ιδέα της αμοιβαίας ενίσχυσης (Mutual Reinforcement), μιας ιδέας που προέρχεται από τη μέθοδο HITS [42] (η τελευταία ήταν βασική ανταγωνίστρια μέθοδος του PageRank και είχε προταθεί για κατάταξη αποτελεσμάτων αναζήτησης σελίδων του διαδικτύου). Οι μέθοδοι που χρησιμοποιούν αυτή τη προσέγγιση εκτελούν υπολογισμούς πάνω σε ετερογενείς διμερής γράφους (Bipartite Networks), όπου οι κόμβοι σε κάθε πλευρά του γράφου ενισχύουν ο ένας τη βαθμολογία του άλλου (π.χ. οι βαθμολογίες των συγγραφέων επηρεάζουν τη βαθμολογία των δημοσιεύσεων και αντίστροφα). Τέτοιοι υπολογισμοί μπορεί να πραγματοποιούνται παράλληλα με υπολογισμούς σε ομογενή δίκτυα (δημοσιεύσεων, συγγραφέων, περιοδικών). Η δεύτερη υποκατηγορία περιλαμβάνει μεθόδους που οργανώνουν όλη τη πληροφορία σε ένα μοναδικό ετερογενή γράφο, πάνω στον οποίο εφαρμόζονται επαναληπτικοί υπολογισμοί.

Στην πρώτη από τις παραπάνω υποκατηγορίες ανήκουν οι μέθοδοι FutureRank [66], P-Rank [85], MR-Rank [93], Wang et al. [80], COIRank [3] και Tri-Rank [52]. Η μέθοδος FutureRank συνδυάζει έναν απλό υπολογισμό PageRank στο δίκτυο αναφορών με υπολογισμούς αμοιβαίας ενίσχυσης μεταξύ δημοσιεύσεων και συγγραφέων. Ταυτόχρονα εισάγει και έναν χρονικό παράγοντα που εξαρτάται από την ηλικία των δημοσιεύσεων (όπως περιγράφεται στην Ενότητα 3.1.2.2). Η μέθοδος P-Rank επίσης χρησιμοποιεί υπολογισμούς PageRank και υπολογισμούς με αμοιβαία ενίσχυση, οι οποίοι γίνονται σε διμερείς γράφους συγγραφέων-δημοσιεύσεων και περιοδικών-δημοσιεύσεων. Η μέθοδος MR-Rank χρησιμοποιεί ένα διμερή γράφο δημοσιεύσεων-περιοδικών. Η μέθοδος αρχικοποιεί τις βαθμολογίες περιοδικών και δημοσιεύσεων στις τιμές που προκύπτουν από υπολογισμούς PageRank στους αντίστοιχους γράφους. Στη συνέχεια εκτελεί υπολογισμούς αμοιβαίας ενίσχυσης μεταξύ περιοδικών και δημοσιεύσεων. Σε αυτούς τους υπολογισμούς, σε κάθε επαναληπτικό βήμα, οι βαθμολογίες των δημοσιεύσεων προκύπτουν από ένα γραμμικό συνδυασμό βαθμολογιών που τους αποδίδονται από τις αναφέρουσες δημοσιεύσεις και από το περιοδικό στο οποίο είναι δημοσιευμένες. Αντίστοιχα οι βαθμολογίες των περιοδικών υπολογίζονται ως γραμμικός συνδυασμός των βαθμολογιών των δημοσιεύσεών τους και των βαθμολογιών άλλων περιοδικών, των οποίων οι δημοσιεύσεις τις αναφέρουν. Στη μέθοδο που παρουσιάζεται από τους Wang et al. χρησιμοποιούνται διμερείς γράφοι δημοσιεύσεων-συγγραφέων και δημοσιεύσεων-περιοδικών. Παράλληλα γίνεται χρήση χρονικών παραμέτρων για συγγραφείς, δημοσιεύσεις και περιοδικά, ενώ παράγονται ταυτόχρονα κατατάξεις δημοσιεύσεων, συγγραφέων και περιοδικών. Η μέθοδος COIRank επεκτείνει το προαναφερθέν μοντέλο τροποποιώντας αναφορές μεταξύ δημοσιεύσεων στη περίπτωση που οι συγγραφείς τους έχουν συνεργαστεί στο παρελθόν, ή όταν εργάζονται στο ίδιο ίδρυμα. Ο σκοπός της μεθόδου είναι να περιορίσει την επίδραση της τεχνητής ενίσχυσης του αριθμού αναφορών που μπορεί να επιδιωχθεί μέσα από κακόβουλες πρακτικές όπως οι αυτοαναφορές (Self Citations), ή οι αμοιβαίες αναφορές (Mutual Citations). Τέλος, η μέθοδος Tri-Rank χρησιμοποιεί διμερείς γράφους δημοσιεύσεων-συγγραφέων, δημοσιεύσεων-περιοδικών και συγγραφέων-περιοδικών. Πραγματοποιεί επαναληπτικούς υπολογισμούς αμοιβαίας

ενίσχυσης για να παράξει βαθμολογίες περιοδικών, συγγραφέων και δημοσιεύσεων. Επιπλέον η μέθοδος χρησιμοποιεί διάφορα βάρη που εφαρμόζονται στις ακμές των διμερών γράφων π.χ. βάσει των αυτοαναφορών, ή βάσει της σειράς με την οποία εμφανίζονται οι συγγραφείς μιας δημοσίευσης.

Η δεύτερη προσέγγιση στη χρήση πολλαπλών δικτύων ακολουθείται από τις μεθόδους PopRank [58] και MutualRank [40]. Η μέθοδος PopRank [58] προσομοιώνει έναν «τυχαίο εντοπιστή αντικειμένων» (Random Object Finder), μια οντότητα που πραγματοποιεί μια τυχαία περιήγηση μεταξύ ιστοσελίδων και «αντικειμένων του διαδικτύου» (Web Objects) που αναπαριστούν δημοσιεύσεις, συγγραφείς και περιοδικά. Οι μεταβάσεις της οντότητας που προσομοιώνεται, οι οποίες πραγματοποιούνται από ένα τύπο αντικειμένου σε ένα άλλο τύπο, εξαρτώνται από κάποιες πιθανότητες που προκύπτουν μετά από εκπαίδευση και ονομάζονται *παράγοντες διάδοσης δημοφιλίας* (Popularity Propagation Factors). Η μέθοδος MutualRank χρησιμοποιεί έναν συνολικό πίνακα γειτνίασης που αποτελείται από 3 υποπίνακες ομογενών δικτύων και 6 υποπίνακες ετερογενών δικτύων. Τα ομογενή δίκτυα αποτελούνται από σταθμισμένες ακμές μεταξύ δημοσιεύσεων, συγγραφέων και περιοδικών/συνεδρίων, ενώ τα ετερογενή δίκτυα αποτελούνται από ακμές μεταξύ των τριών προαναφερθέντων υποδικτύων. Η μέθοδος παράγει ταυτόχρονα μια κατάταξη όλων των κόμβων (που αντιστοιχούν σε δημοσιεύσεις, συγγραφείς, περιοδικά) με βάση έναν υπολογισμό ιδιοδιανύσματος πάνω στο συνολικό πίνακα γειτνίασης.

### 3.1.5 Συνδυαστικές Μέθοδοι

Οι συνδυαστικές μέθοδοι υλοποιούν πολλαπλές μεθόδους κατάταξης και συνδυάζουν τις βαθμολογίες που προκύπτουν για να παράξουν μια τελική βαθμολογία ανά δημοσίευση. Η πλειοψηφία των μεθόδων που προτάθηκαν στο διαγωνισμό 2016 WSDM Cup[1] ανήκουν σ' αυτή τη κατηγορία. Ο διαγωνισμός είχε ως αντικείμενο την κατάταξη δημοσιεύσεων βάσει της «σημασίας ανεξαρτήτως ερωτήματος» (Query-independent Importance), με χρήση πολλαπλών διασυνδεδεμένων δικτύων [77].

Η μέθοδος NTUTriPartite [22] που κέρδισε το διαγωνισμό, χρησιμοποιεί έναν γραμμικό συνδυασμό του αρχικού βαθμού εισερχόμενων και εξερχόμενων ακμών των κόμβων και υπολογισμούς διάδοσης βαθμολογιών μεταξύ διαφόρων δικτύων. Οι υπολογισμοί πραγματοποιούνται επαναληπτικά, με προκαθορισμένο το πλήθος των επαναλήψεων. Η μέθοδος NTUEnsemble [14] συνδυάζει βαθμολογίες απο τη μέθοδο NTUWeightedPR [16] (Ενότητα 3.1.3) με βαθμολογίες από τη νικητήρια μέθοδο του διαγωνισμού και επιπλέον βαθμολογίες που βασίζονται στη μέθοδο των Wang et al. [80] (Ενότητα 3.1.4). Η μέθοδος EWPR [53], από την άλλη, χρησιμοποιεί ένα συνδυασμό βαθμολογιών περιοδικών και δημοσιεύσεων που έχουν προκύψει από υπολογισμούς PageRank με χρονικούς παράγοντες και βαθμολογιών συγγραφέων που υπολογίζονται από το μέσο όρο των βαθμολογιών PageRank των δημοσιεύσεών τους. Η μέθοδος SARank [55] αποτελεί επέκταση της παραπάνω μεθόδου, συμπεριλαμβάνοντας επιπλέον μια βαθμολογία που βασίζεται σε εκθετικά σταθμισμένους αριθμούς αναφορών, καθώς και επιπλέον βαθμολογίες συγγραφέων και δημοσιεύσεων βασισμένες στους μέσους όρους των παραπάνω σταθμισμένων αριθμών αναφορών. Στη μέθοδο ALEF [82] γίνεται χρήση του Article Level Eigenfactor (μιας μεθόδου που βασίζεται στο PageRank) για τον υπολογισμό των βαθμολογίων των δημοσιεύσεων. Βάσει αυτών των βαθμολογιών υπολογίζονται βαθμολογίες για συγγραφείς, οι οποίες συνδυάζονται με τις

[1] http://www.wsdm-conference.org/2016/wsdm-cup.html

βαθμολογίες δημοσιεύσεων ως σταθμισμένο άθροισμα. Τέλος, η μέθοδος bletchley-park [35] υπολογίζει τις βαθμολογίες κάθε δημοσίευσης ως ένα γραμμικό συνδυασμό του αριθμού αναφορών, των βαθμολογιών PageRank, της ηλικίας των δημοσιεύσεων, καθώς και κάποιων βαθμολογιών συγγραφέων και περιοδικών που προκύπτουν από το σύνολο των δημοσιεύσεών τους.

### 3.1.6 Άλλες Μέθοδοι

Σε αυτή την ενότητα εξετάζουμε ορισμένες μεθόδους της βιβλιογραφίας που δεν εμπίπτουν στην παραπάνω κατηγοριοποίηση.

Η μέθοδος S-RCR [64] αποτελεί μια απλοποίηση της μεθόδου RCR [36] (η τελευταία βασίζεται σε μηχανική μάθηση). Η μέθοδος χρησιμοποιεί το ρυθμό αναφορών μιας δημοσίευσης, δηλαδή το λόγο του αριθμού αναφορών προς την ηλικία της, για να υπολογίσει τη βαθμολογία μιας δημοσίευσης. Η ρυθμός αναφορών συγκρίνεται με αυτόν όλων των άλλων δημοσιεύσεων που ανήκουν στην ίδια «γειτονιά». Ο όρος «γειτονιά» αναφέρεται σε όλες τις δημοσιεύσεις $j$ που εμφανίζονται μαζί με τη δημοσίευση $i$ (η οποία βαθμολογείται) σε κάποια λίστα αναφορών. Η μέθοδος Citation Wake [43] υπολογίζει τις βαθμολογίες των δημοσιεύσεων ως σταθμισμένο άθροισμα όλων των δημοσιεύσεων που απέχουν κατά συντομότερα μονοπάτια $\lambda$ βήματα από τη δημοσίευση που βαθμολογείται. Σημειώνουμε ότι κάθε δημοσίευση $j$ που αναφέρει τη δημοσίευση $i$, συμπεριλαμβάνεται μόνο μια φορά στους υπολογισμούς (με βάση τη μικρότερη απόσταση). Ως εκ τούτου η μέθοδος *δεν* χρησιμοποιεί όλα τα διαφορετικά μονοπάτια του γράφου, αλλά τα μεγέθη των συνόλων των δημοσιεύσεων που βρίσκονται σε ελάχιστη απόσταση $\lambda$ αναφορών από τη βαθμολογούμενη. Η μέθοδοι Age-Rescaled PageRank [56] και Age- and Field-Rescaled PageRank [74] πραγματοποιούν έναν αρχικό υπολογισμό PageRank του οποίου τα αποτελέσματα κανονικοποιούνται. Η κανονικοποίηση που γίνεται βασίζεται στους αριθμητικούς μέσους και την τυπική απόκλιση των βαθμολογιών των $n$ δημοσιεύσεων που έχουν δημοσιευτεί πλησιέστερα χρονικά πριν και μετά από κάθε υπό εξέταση δημοσίευση. Στην περίπτωση της μεθόδου Age- and Field-Rescaled PageRank αυτή η κανονικοποίηση επιπλέον έχει τον περιορισμό ότι λαμβάνει υπόψη μόνο δημοσιεύσεις από το ίδιο επιστημονικό πεδίο. Τέλος στη μέθοδο που προτείνεται από τους Bai et al. [1] χρησιμοποιείται μια παραλλαγή του Quantum PageRank [60], όπου τα βάρη των ακμών αναφορών προσδιορίζονται από μια συνάρτηση της γεωγραφικής απόστασης των ιδρυμάτων από τα οποία έχουν προκύψει οι δύο δημοσιεύσεις.

Πίνακας 3.1: Κατηγοριοποίηση Μεθόδων Κατάταξης Δημοσιεύσεων. Για τις μεθόδους που αξιολογούνται πειραματικά στο Κεφάλαιο 4 χρησιμοποιείται πιο έντονη γραμματοσειρά.

| **Μέθοδος** | **Απλές** | **Χρονικές Παράμετροι** | | | **Μεταδεδομένα** | | **Πολλαπλά** | **Συνδυαστικές** | **Άλλες** |
|---|---|---|---|---|---|---|---|---|---|
| | **Παραλλαγές PR** | **Πίνακας Δικτύου** | | **Πιθανότητα Επιλογής** | **Χώρος Δημοσίευσης** | **Συγγραφείς** | **Δίκτυα** | | |
| **Non-Linear PageRank (NPR) [87]** | ✓ | | | | | | | | |
| SPR [95] | ✓ | | | | | | | | |
| SCEAS [67] | ✓ | | | | | | | | |
| Focused PageRank [46] | ✓ | | | | | | | | |
| PrestigeRank [72] | ✓ | | | | | | | | |
| **Weighted Citation (WC) [84]** | | ✓ | | | ✓ | | | | |
| **Retained Adjacency Matrix (RAM) [28]** | | ✓ | | | | | | | |
| Timed PageRank [89, 90] | | ✓ | | | ✓ | ✓ | | | |
| **Effective Contagion Matrix (ECM) [28]** | | ✓ | | | | | | | |
| **NewRank (NR) [18]** | | ✓ | | ✓ | | | | | |
| NTUWeightedPR [16] | | ✓ | | ✓ | ✓ | ✓ | | | |
| EWPR [53] | | ✓ | | | ✓ | ✓ | | ✓ | |
| SARank [55] | | ✓ | | | ✓ | ✓ | | ✓ | |
| **CiteRank (CR) [78]** | | | | ✓ | | | | | |
| **FutureRank (FR) [66]** | | | | ✓ | | ✓ | ✓ | | |
| MR-Rank [93] | | ✓ | | | ✓ | | ✓ | | |
| P-Rank [85] | | | | | ✓ | ✓ | ✓ | | |
| **YetRank (YR) [37]** | | | | ✓ | ✓ | | | | |
| Wang et al. [80] | | | | ✓ | ✓ | ✓ | ✓ | | |
| COIRank. [3] | | | | ✓ | ✓ | ✓ | ✓ | | |
| PopRank [58] | | | | | | | ✓ | | |
| MutualRank [40] | | | | | | | ✓ | | |
| Tri-Rank [52] | | | | | ✓ | ✓ | ✓ | | |
| **NTUTriPartite (WSDM) [22]** | | | | | ✓ | ✓ | ✓ | ✓ | |
| NTUEnsemble [14] | | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | |
| bletchleypark [35] | | ✓ | | | ✓ | ✓ | | ✓ | |
| ALEF [82] | | | | | | ✓ | | ✓ | |
| S-RCR [64] | | | | | | | | | ✓ |
| Citation Wake [43] | | | | | | | | | ✓ |
| Age-Rescaled PR [56] | | | | | | | | | ✓ |
| Age- & Field- Rescaled PR [74] | | | | | | | | | ✓ |
| Bai et al. [1] | | | | | | | | | ✓ |

## 3.2 Κατηγοριοποίηση Τρόπων Αξιολόγησης Μεθόδων Κατάταξης Δημοσιεύσεων στη Βιβλιογραφία

Οι μέθοδοι κατάταξης επιστημονικών δημοσιεύσεων που εξετάστηκαν στην Ενότητα 3.1 προέρχονται από διαφορετικά επιστημονικά πεδία. Συνεπώς, στη βιβλιογραφία δεν υπάρχει κάποιος ενιαίος στόχος όσο αφορά την κατάταξη, ενώ τα κριτήρια του τι καθιστά μια κατάταξη αποτελεσματική μπορεί να μην είναι πάντοτε σαφώς δοσμένα. Επιπλέον οι διάφορες μέθοδοι έχουν αξιολογηθεί σε διαφορετικά σύνολα δεδομένων με χρήση διαφορετικών θεωρήσεων και μετρικών αξιολόγησης. Σκοπός της παρούσας ενότητας είναι να εξετάσει τους διάφορους στόχους και να κατηγοριοποιήσει τις διάφορες μεθοδολογίες που εφαρμόζονται για την αξιολόγηση των μεθόδων κατάταξης στη βιβλιογραφία.

### 3.2.1 Αξιολόγηση με Κριτήριο την Ποιότητα της Κατάταξης

**Υπόβαθρα Αληθείας (Ground Truth Lists)**. Ένας τρόπος αξιολόγησης της αποτελεσματικότητας μιας μεθόδου κατάταξης δημοσιεύσεων είναι η χρήση μιας λίστας κατάταξης που αποτελεί το υπόβαθρο αληθείας. Οι λίστες αυτές δίνουν είτε κάποια συνολική, είτε κάποια μερική κατάταξη (δηλαδή μόνο των θεωρούμενων ως πιο σημαντικών) δημοσιεύσεων. Η αξιολόγηση των μεθόδων γίνεται με τη σύγκριση της κατάταξης που παράγουν σε σχέση με την κατάταξη στη λίστα αναφοράς, η οποία θεωρείται ως η ιδανική. Συνήθως, όταν χρησιμοποιείται κάποιο υπόβαθρο αληθείας, ο σκοπός είναι ο εντοπισμός των δημοσιεύσεων που έχουν μεγάλη μακροχρόνια απήχηση. Συνεπώς το υπόβαθρο αληθείας αποτελείται από βραβευμένες δημοσιεύσεις, ή από δημοσιεύσεις που έχουν επιλεγεί από ειδικούς ως οι πιο σημαντικές. Τέτοιου τύπου αξιολόγηση ακολουθείται στα [87, 95, 18, 37, 56, 85]. Αντίστοιχα στο [43] το υπόβαθρο αληθείας βασίζεται σε δημοσιεύσεις που έχουν συγγραφεί από βραβευμένους επιστήμονες. Η χρήση τέτοιων υποβάθρων αληθείας για την αξιολόγηση μεθόδων κατάταξης έχει ορισμένα εγγενή μειονεκτήματα: μπορεί να μη παρέχουν μια συνολική κατάταξη, να είναι μεροληπτικά (όταν προσδιορίζονται από ομάδα ειδικών), ή μπορεί να μην είναι διαθέσιμα για όλους τους επιστημονικούς κλάδους.

**Κρίσεις Χρηστών (User Judgements)**. Ένας άλλος τρόπος αξιολόγησης της αποτελεσματικότητας των μεθόδων κατάταξης είναι η χρήση κρίσεων χρηστών, όπως γίνεται στα [58, 55] και το διαγωνισμό 2016 WSDM Cup. Σε αυτό το είδος αξιολόγησης παρέχεται από ειδικούς η σχετική αξία ανάμεσα σε ζευγάρια δημοσιεύσεων. Τα αποτελέσματα κατάταξης που παράγονται από τις μεθόδους κατάταξης συγκρίνονται στη συνέχεια με αυτές τις σχετικές κατατάξεις.

**Παρακρατημένα Δεδομένα (Held-Out Data.)** Ένας δημοφιλής τρόπος αξιολόγησης που εφαρμόζεται στα [87, 78, 66, 28, 89, 90, 95, 55, 93] είναι η εκτίμηση της ακρίβειας με την οποία κάθε μέθοδος παράγει μια κατάταξη βασισμένη στα είδη απήχησης που περιγράφηκαν στην Ενότητα 2.4. Καθώς αυτά τα μέτρα απήχησης καθορίζονται από μια μελλοντική κατάσταση του δικτύου αναφορών, η διαδικασία αξιολόγησης περιλαμβάνει την «παρακράτηση» τμήματος του συνόλου δεδομένων. Συγκεκριμένα επιλέγεται κάποια χρονική στιγμή $t_c$ στο σύνολο δεδομένων η οποία αναπαριστά μια

εικονική τρέχουσα στιγμή. Δημιουργείται έτσι μια *τρέχουσα κατάσταση* του δικτύου αναφορών που περιγράφεται από τον πίνακα γειτνίασης $\boldsymbol{A}(t_c)$ και μια *μελλοντική κατάσταση* που περιγράφεται από τον πίνακα γειτνίασης $\boldsymbol{A}(t_c + T)$, όπου η παράμετρος $T$ ορίζει κάποιον χρονικό ορίζοντα.

Η αξιολόγηση γίνεται ακολουθώντας τα εξής βήματα: πρώτον, παράγεται μια κατάταξη αναφοράς, βάσει του είδους απήχησης που ενδιαφέρει. Αυτή χρησιμοποιείται ως υπόβαθρο αληθείας. Δεύτερον, εφαρμόζεται η υπό αξιολόγηση μέθοδος κατάταξης στο τμήμα του δικτύου που περιγράφεται από τον πίνακα $\boldsymbol{A}(t_c)$. Στο τρίτο βήμα, μετράται η συμφωνία της κατάταξης που έχει παραχθεί από τη μέθοδο με την κατάταξη βάση του μέτρου απήχησης. Στη βιβλιογραφία, η παραπάνω μέθοδος αξιολόγησης έχει χρησιμοποιηθεί για την αποτελεσματικότητα κατάταξης βάσει της βραχυχρόνιας απήχησης (ή δημοφιλίας), μετρούμενης με χρήση του αριθμού αναφορών, στα [78, 28, 89, 90]. Αναφερόμαστε σε αυτή τη περίπτωση ως P-CC (Popularity-Citation Count). Αντίστοιχα στο [66] χρησιμοποιείται ως μέτρο της δημοφιλίας το PageRank, πράγμα που συμβολίζουμε ως P-PR (Popularity-PageRank). Στο [28] γίνεται επιπλέον αξιολόγηση με βάση τη μακροχρόνια απήχηση (ή συνολική επιρροή), με χρήση του PageRank ως μέτρου απήχησης. Αναφερόμαστε στην περίπτωση αυτή ως I-PR (Influence-PageRank). Έως τώρα στη βιβλιογραφία δεν έχουμε εντοπίσει χρήση του αριθμού αναφορών σαν μέτρο μακροχρόνιας απήχησης, κάτι το οποίο αναφέρουμε ως I-CC (Influence-Citation Count).

Ένα σημαντικό σημείο στην παραπάνω διαδικασία αξιολόγησης αφορά την επιλογή του χρονικού σημείου $t_c$. Με τη μακροχρόνια απήχηση μετράμε κάποιο μέτρο κεντρικότητας στο δίκτυο αναφορών άπειρες χρονικές μονάδες στο μέλλον. Αφενός για να εκτιμήσουμε σωστά τη μακροχρόνια απήχηση απαιτείται μια μελλοντική κατάσταση του δικτύου που αφορά το μακρινό μέλλον, επομένως όσον αφορά το σύνολο δεδομένων στο οποίο γίνεται η αξιολόγηση, η τιμή του $t_c$ πρέπει να είναι μικρή. Από την άλλη, για να μπορούμε να αξιολογήσουμε ορθά μια μέθοδο κατάταξης, απαιτείται το σύνολο στο οποίο τρέχει να περιέχει επαρκή δεδομένα, πράγμα που απαιτεί ένα επαρκώς μεγάλο $t_c$. Είναι φανερό ότι οι απαιτήσεις αυτές είναι σε αντίφαση. Αντίστοιχα στην περίπτωση της βραχυχρόνιας απήχησης, το $t_c$ επιλέγεται σχετικά κοντά στην τελική χρονική στιγμή του συνόλου δεδομένων. Γενικά, οι επιπτώσεις που έχει η εκάστοτε επιλογή του σημείου $t_c$ δεν έχουν μελετηθεί. Έως τώρα οι σχετικές εργασίες το ορίζουν είτε σε κάποιο συγκεκριμένο έτος, είτε έτσι ώστε το σύνολο δεδομένων να χωρίζεται με τρόπο τέτοιο ώστε να ορίζεται συγκεκριμένος λόγος του αριθμού δημοσιεύσεων που περιέχονται στην μέλλουσα προς την τρέχουσα κατάσταση.

### 3.2.2 Αξιολόγηση με Κριτήρια μη Σχετικά με την Ποιότητα Κατάταξης

**Περιγραφική Αξιολόγηση (Descriptive Evaluation)**. Κάποιες μέθοδοι της βιβλιογραφίας αξιολογούνται έμμεσα στις εργασίες όπου παρουσιάζονται, ενώ δεν ορίζεται κάποιος συγκεκριμένος στόχος κατάταξης. Σε κάποιες περιπτώσεις περιγράφονται οι δημοσιεύσεις που έρχονται πρώτες σε κατάταξη. Παρουσιάζεται κάποια επιχειρηματολογία γύρω από το γιατί η παραγόμενη κατάταξη έχει ενδιαφέρον, ή κάποια ιδιαίτερη αξία, προσέγγιση που ακολουθείται στα [66, 18, 89, 90, 54, 15, 43, 72]. Σε άλλες περιπτώσεις παρουσιάζεται η σχέση της παραγόμενης κατάταξης με αυτή κάποιας άλλης μεθόδου αναφοράς (π.χ. τον αριθμό αναφορών, το PageRank, ή κάποια άλλη μέθοδο). Συνήθως για το λόγο αυτό χρησιμοποιούνται διαγράμματα διασποράς (Scatter

Plots), όπως στα [15, 54, 78, 84, 18], και τιμές συσχέτισης των κατατάξεων, όπως στα [54, 95, 93, 72, 52]. Μια άλλη προσέγγιση, η οποία ακολουθείται στα [15, 37] είναι η παρουσίαση της μέσης βαθμολογίας κατάταξης των δημοσιεύσεων συναρτήσει του έτους δημοσίευσής τους. Αυτού του τύπου η αξιολόγηση είναι χρήσιμη ως απόδειξη ότι μια μέθοδος δεν είναι μεροληπτική ενάντια σε πρόσφατες δημοσιεύσεις.

**Ανεξάρτητη από την Αποτελεσματικότητα Αξιολόγηση.** Κάποιες εργασίες εστιάζουν στην αξιολόγηση πτυχών των μεθόδων κατάταξης, που δεν σχετίζονται με την αποτελεσματικότητα των τελευταίων. Για παράδειγμα, στα [66, 78] παρουσιάζονται οι χρόνοι εκτέλεσης των μεθόδων κατάταξης, ενώ στα [66, 95, 93, 52] εξετάζεται η ταχύτητα σύγκλισης τους. Στα [95, 87] η αξιολόγηση βασίζεται στο πόσο σταθερά είναι τα αποτελέσματα της κατάταξης απέναντι σε κακόβουλες πρακτικές, όπως για παράδειγμα απέναντι σε αυτοαναφορές. Τέλος, σε εργασίες όπως οι [43, 89, 90] εξετάζεται η σταθερότητα των αποτελεσμάτων κατάταξης όταν μεταβάλλονται οι διάφορες ελεύθερες παράμετροι της μεθόδου.

# Κεφάλαιο 4

# Πειραματική Αξιολόγηση Τεχνολογιών Αιχμής Κατάταξης Δημοσιεύσεων

Σε αυτό το κεφάλαιο πραγματοποιούμε πειράματα για την αξιολόγηση της αποτελεσματικότητας των μεθόδων κατάταξης δημοσιεύσεων. Συγκεκριμένα, στην Ενότητα 4.1 θέτουμε τρία ερευνητικά ερωτήματα γύρω από τα οποία αναπτύσσουμε τη πειραματική μας μελέτη, περιγράφουμε την μεθοδολογία αξιολόγησης που ακολουθούμε και τις μεθόδους κατάταξης που εξετάζουμε. Έπειτα, στις Ενότητες 4.2-4.4 πραγματοποιούμε σειρά πειραμάτων βασιζόμενοι στη μεθοδολογία αυτή και, τέλος, παρουσιάζουμε τα συμπεράσματά μας στην Ενότητα 4.5.

## 4.1 Πλαίσιο Αξιολόγησης

Εισαγωγικά, θέτουμε τρία ερευνητικά ερωτήματα, γύρω από τα οποία θα αναπτύξουμε τη μελέτη αυτού του κεφαλαίου. Έπειτα, περιγράφουμε τα σύνολα δεδομένων πάνω στα οποία εκτελούμε τα πειράματά μας, τις μεθόδους κατάταξης δημοσιεύσεων που εξετάζουμε και τις μετρικές αξιολόγησης που χρησιμοποιούμε για να απαντήσουμε στα ερωτήματα αυτά.

### 4.1.1 Ερευνητικά Ερωτήματα

Τα ερευνητικά ερωτήματα που μας απασχολούν είναι τα ακόλουθα:

- **Ερευνητικό Ερώτημα 1: Πόσο διακριτές είναι οι έννοιες της βραχυχρόνιας και μακροχρόνιας απήχησης (ή της δημοφιλίας και επιρροής, αντίστοιχα);** Στην Ενότητα 4.2 διερευνούμε τη σχέση μεταξύ των κατατάξεων που χρησιμοποιούμε ως υπόβαθρα αληθείας, δηλαδή των I-CC, I-PR, P-CC, P-PR.
- **Ερευνητικό Ερώτημα 2: Ποιες είναι οι πιο αποτελεσματικές μέθοδοι στην παραγωγή κατατάξεων για κάθε είδος απήχησης;** Αξιολογούμε την αποτελεσματικότητα μεθόδων που αποτελούν τεχνολογίες αιχμής στην κατάταξη δημοσιεύσεων. Συγκεκριμένα, στις Ενότητες 4.3.1 και 4.3.2 εξετάζουμε την αποτελεσματικότητα ως προς την παραγωγή κατατάξεων βάσει της μακροχρόνιας και βραχυχρόνιας απήχησης, αντίστοιχα.

- **Ερευνητικό Ερώτημα 3: Πόσο γρήγορα συγκλίνουν οι επαναληπτικές μέθοδοι;** Δεδομένου ότι οι περισσότερες μέθοδοι της βιβλιογραφίας βασίζονται σε επαναληπτικούς αλγορίθμους, οι οποίοι εκτελούνται έως ότου ικανοποιηθεί κάποιο κριτήριο σύγκλισης, συγκρίνουμε τις μεθόδους ως προς την ταχύτητα σύγκλισης (και συνεπώς τον χρόνο εκτέλεσης) για να ανακαλύψουμε τα σχετικά μειονεκτήματα, ή πλεονεκτήματα στη χρήση κάθε μιας, έναντι κάποιας άλλης. Έτσι μπορούμε να επιλέξουμε την πιο κατάλληλη ανάμεσα σε μεθόδους που έχουν συγκρίσιμη αποτελεσματικότητα, σε περιπτώσεις όπου απαιτείται η παραγωγή κατατάξεων υπό χρονικούς περιορισμούς. Εξετάζουμε την ταχύτητα σύγκλισης και τους χρόνους εκτέλεσης των επαναληπτικών μεθόδων στην Ενότητα 4.4.[1]

### 4.1.2 Σύνολα Δεδομένων

Στα πειράματά μας χρησιμοποιούμε τα ακόλουθα σύνολα δεδομένων:

- *hep-th*[2]: αποτελείται από περίπου 30,000 δημοσιεύσεις στο αντικείμενο της θεωρητικής φυσικής υψηλής ενέργειας (high enegry physics - theory) από το ηλεκτρονικό αρχείο arXiv, οι οποίες δημοσιεύτηκαν από το 1992 έως το 2003.

- *APS*[3]: αποτελείται από περίπου μισό εκατομμύριο δημοσιεύσεις από τα περιοδικά της αμερικάνικης εταιρίας φυσικής (American Physical Society), τα οποία δημοσιεύτηκαν από το 1893 έως το 2014.

- *PMC*[4]: αποτελείται από περίπου 1.12 εκατομμύρια ελεύθερα διαθέσιμες δημοσιεύσεις από το χώρο των επιστημών ζωής (Life Sciences), δημοσιευμένες από το 1896 έως το 2016.

- *DBLP*[5]: αποτελείται από περίπου 3 εκατομμύρια δημοσιεύσεις από το χώρο της πληροφορικής, οι οποίες καταγράφονται στο DBLP, δημοσιευμένες από το 1936 έως το 2018.

  Σημειώνουμε ότι τα πρώτα δύο σύνολα δεδομένων έχουν χρησιμοποιηθεί ευρέως σε παλαιότερες εργασίες (π.χ., [15, 66, 78, 28, 87]), ενώ τα τελευταία δύο είναι αντιπροσωπευτικά πραγματικών μεγάλων συνόλων δημοσιεύσεων (με περισσότερους από 1 εκατομμύριο κόμβους η κάθε μία), από δύο σημαντικούς και παραγωγικούς κλάδους της επιστήμης: τις βιοεπιστήμες και τη πληροφορική.

### 4.1.3 Υλοποιήσεις Μεθόδων

Οι μέθοδοι που αξιολογούνται πειραματικά σε αυτό το κεφάλαιο, επιλέχθηκαν με κριτήριο να καλύψουμε όλες τις διαφορετικές κατηγορίες μεθόδων που περιγράψαμε στην

[1] Εξαιρούνται από τη διαδικασία, εκτός από τις μεθόδους που βασίζονται σε υπολογισμό του αριθμού αναφορών (δεδομένου ότι δεν είναι επαναληπτικές), οι μέθοδοι που εκτελούνται με βάση ένα προκαθορισμένο αριθμό επαναλήψεων.

[2] http://www.cs.cornell.edu/projects/kddcup/datasets.html

[3] http://journals.aps.org/about

[4] ftp://ftp.ncbi.nlm.nih.gov/pub/pmc

[5] https://aminer.org/citation

Ενότητα 3.1.

**PageRank (PR).** Υλοποιούμε τον κλασσικό αλγόριθμο που περιγράφεται στο [59] (Ενότητα 2.2.2).

**Non-Linear PageRank (NPR).** Πρόκειται για μια απλή παραλλαγή του PageRank που περιγράφεται στο [87] (Ενότητα 3.1).

**CiteRank (CR).** Πρόκειται για παραλλαγή του PageRank που χρησιμοποιεί πιθανότητες επιλογής εκθετικά φθίνουσες με την ηλικία των δημοσιεύσεων [78] (Ενότητα 3.1.2).

**FutureRank (FR).** Αυτή η παραλλαγή του PageRank χρησιμοποιεί χρονικά φθίνουσες με την ηλικία των δημοσιεύσεων πιθανότητες επιλογής, καθώς και πολλαπλά δίκτυα [66]. Η βασική του ιδέα είναι το μοίρασμα των βαθμολογιών των δημοσιεύσεων στους συγγραφείς τους και αντίστροφα, μια ιδέα εμπνευσμένη από τη μέθοδο HITS [42].

**Retained Adjacency Matrix (RAM).** Αυτή η μέθοδος αποτελεί τροποποίηση του αριθμού αναφορών, όπου κάθε ακμή έχει βάρος που φθίνει εκθετικά με την ηλικία της αναφοράς [28].

**Effective Contagion Matrix (ECM).** Αποτελεί παραλλαγή του μέτρου κεντρικότητας Katz και χρησιμοποιεί χρονικά ενήμερο πίνακα γειτνίασης, με χρήση της ηλικίας αναφοράς [28].

**NewRank (NR).** Αυτή η παραλλαγή του PageRank χρησιμοποιεί ένα χρονικά ενήμερο πίνακα γειτνίασης με βάση την ηλικία της αναφερόμενης δημοσίευσης, καθώς και πιθανότητες επιλογής που φθίνουν εκθετικά με την ηλικία των δημοσιεύσεων [18].

**YetRank (YR).** Αυτή η παραλλαγή του PageRank χρησιμοποιεί πιθανότητες επιλογής εκθετικά φθίνουσες με την ηλικία των δημοσιεύσεων, οι οποίες επιπλέον καθορίζονται από τον Impact Factor του περιοδικού (υπολογισμένο με χρήση των δεδομένων της τελευταίας πενταετίας) στο οποίο εκδόθηκε η κάθε δημοσίευση [37].

**NTUTriPartite (WSDM).** Πρόκειται για συνδυαστική μέθοδο που δε χρησιμοποιεί χρονικές παραμέτρους. Υπολογίζει τις βαθμολογίες κάθε δημοσίευσης βάσει των συγγραφέων της, του περιοδικού όπου δημοσιεύτηκε, των αναφερουσών δημοσιεύσεων, καθώς και ενός συνδυασμού του αριθμού αναφορών και του μεγέθους της λίστας αναφορών κάθε δημοσίευσης. Η μέθοδος είναι επαναληπτική και εκτελείται με βάση έναν προκαθορισμένο αριθμό επαναλήψεων [22].

**Weighted Citation (WC)**. Η μέθοδος αυτή αποτελεί παραλλαγή του αριθμού αναφορών [84], η οποία χρησιμοποιεί ένα χρονικά ενήμερο πίνακα γειτνίασης με χρονική ποσότητα το χρονικό διάστημα αναφοράς. Επιπλέον τα βάρη των ακμών εξαρτώνται από την τιμή Eigenfactor [5] του περιοδικού της αναφέρουσας δημοσίευσης. Η τιμή αυτή αποτελεί ένα μέτρο παρόμοιο με τον Impact Factor των περιοδικών. Συγκεκριμένα, ελλείψει των τιμών του Eigenfactor, στην υλοποίηση της μεθόδου έγινε χρήση του Impact Factor, υπολογισμένου στα δεδομένα της τελευταίας πενταετίας.

Σημειώνουμε ότι οι τελευταίες τρεις μέθοδοι (YR, WSDM, WC) προϋποθέτουν χρήση δεδομένων για τα περιοδικά στα οποία εκδόθηκαν οι δημοσιεύσεις. Τέτοια δεδομένα υπήρχαν διαθέσιμα μόνο για τα σύνολα δεδομένων PMC και DBLP και ως εκ τούτου, οι μέθοδοι αυτές αξιολογήθηκαν μόνο σε αυτά. Στα πειράματα αυτού του κεφαλαίου έχουμε χρησιμοποιήσει για κάθε μέθοδο τις τιμές των παραμέτρων που προτάθηκαν ως βέλτιστες από τους συγγραφείς της εκάστοτε εργασίας. Από τα πειραματικά δεδομένα που συλλέξαμε απεικονίζουμε στις γραφικές παραστάσεις μόνο τα

αποτελέσματα της παραμετροποίησης που οδηγεί στη μεγαλύτερη αποτελεσματικότητα για κάθε πείραμα.

Όλες οι μέθοδοι που εξετάσαμε πειραματικά έχουν υλοποιηθεί σε Python 2.7 και έχουν γίνει ελεύθερα διαθέσιμες με άδεια χρήσης GNU GPL.[6] Όλα τα πειράματα εκτελέστηκαν σε μια συστάδα από 10 εικονικά μηχανήματα (VMs) με 4 πυρήνες και 8 GB μνήμη το καθένα. Τα μηχανήματα αυτά παρείχε η υπηρεσία νέφους ~okeanos [44]. Στην περίπτωση των επαναληπτικών μεθόδων που βασίζονται στη σύγκλιση, τέθηκε ως σφάλμα σύγκλισης η τιμή $10^{-12}$, ώστε να εξασφαλιστεί ότι οι τελικές κατατάξεις που παράγονται δεν θα μεταβάλλονταν με επιπλέον επαναλήψεις (δηλαδή οι τιμές των βαθμολογιών όλων των δημοσιεύσεων πρακτικά δε διαφέρουν από τις ιδανικές).

### 4.1.4 Μέθοδος και Μετρικές Αξιολόγησης

Για να απαντήσουμε τα ερευνητικά ερωτήματα που θέσαμε, διεξάγουμε πειράματα με τη προσέγγιση των παρακρατημένων δεδομένων, όπως περιγράφεται στην Ενότητα 3.2.1. Συγκεκριμένα, για κάθε σύνολο δεδομένων θέτουμε την τιμή του $t_c$, έτσι ώστε η *τρέχουσα κατάσταση*, η οποία περιγράφεται από τον πίνακα $\boldsymbol{A}(t_c)$, να περιέχει τις μισές δημοσιεύσεις σε σχέση με το σύνολο των δεδομένων. Καθορίζουμε αντίστοιχα τη *μελλοντική κατάσταση*, που περιγράφεται από τον πίνακα $\boldsymbol{A}(t_c + T)$, επιλέγοντας το $T$ έτσι ώστε ο λόγος, $\eta$, του αριθμού των δημοσιεύσεων στην τρέχουσα και στη μελλοντική κατάσταση να παίρνει τις τιμές $\{1.2, 1.4, 1.6, 1.8, 2.0\}$. Σαν τιμή βάσης (Default) για το λόγο αυτό θέτουμε το $\eta = 1.6$. Κατασκευάζουμε έπειτα με χρήση της τρέχουσας και μέλλουσας κατάστασης το εκάστοτε υπόβαθρο αληθείας, ανάλογα με το είδος απήχησης που θέλουμε να εξετάσουμε (βασισμένο δηλαδή στον αριθμό αναφορών, ή το PageRank), όπως περιγράφουμε στην Ενότητα 3.2.1.

Για να απαντήσουμε στο πρώτο ερευνητικό ερώτημα που θέσαμε, μετράμε τη συσχέτιση μεταξύ των κατατάξεων που χρησιμοποιούμε ως υπόβαθρα αληθείας. Για να μετρήσουμε τη συσχέτιση χρησιμοποιούμε τα μέτρα $\rho$ του Spearman (Spearman's $\rho$) και $\tau$ του Kendall (Kendall's $\tau$). Για να απαντήσουμε το δεύτερο ερευνητικό ερώτημα, μετράμε τη συσχέτιση μεταξύ της κατάταξης που παράγει κάθε μέθοδος σε σχέση με το κάθε υπόβαθρο αληθείας, χρήση των Spearman's $\rho$ και Kendall's $\tau$, καθώς και την αποτελεσματικότητα κατάταξης ως προς τα $k$-σημαντικότερα (Top-$k$) αποτελέσματα, χρήσει της ακρίβειας και του μέτρου nDCG.

Το $\rho$ του Spearman είναι συνάρτηση της $L1$ νόρμας των θέσεων κατάταξης όλων των στοιχείων σε δύο λίστες κατάταξης [71]. Το $\tau$ του Kendall υπολογίζεται βάσει του αριθμού των ζευγών στοιχείων που κατατάσσονται με την ίδια σειρά μεταξύ δύο λιστών [41]. Η ακρίβεια στα $k$-σημαντικότερα αποτελέσματα ορίζεται ως το ποσοστό των κοινών στοιχείων που κατατάσσονται στις $k$ πρώτες θέσεις σε κάθε λίστα. Το μέτρο DCG στη θέση κατάταξης $k$, στο οποίο βασίζεται το μέτρο nDCG, υπολογίζεται ως $\text{DCG@}k = \sum_{i=1}^{k} \frac{rel(i)}{\log_2(i+1)}$, όπου $rel(i)$ είναι η βαθμολογία που έχει βάσει του υπόβαθρου αληθείας (I-CC, I-PR, P-CC, ή P-PR) η δημοσίευση που βρίσκεται στην $i$-οστή θέση της κατάταξης που παράγει μια μέθοδος. Το μέτρο nDCG στη θέση κατάταξης $k$, αποτελεί κανονικοποιημένη μορφή του παραπάνω, όπου nDCG@$k$ είναι ο λόγος της τιμής DCG της δημοσίευσης, προς την ιδανική τιμή DCG, την οποία θα πετυχαίναμε όταν η κατάταξη μιας μεθόδου συμπίπτει με αυτή του υπόβαθρου αληθείας. Οι δύο τελευταίες μετρικές που αναφέραμε, υπολογίζονται με αναφορά σε κάποια θέση κατάταξης

[6] https://github.com/diwis/PaperRanking

Πίνακας 4.1: Συσχετίσεις (Spearman's $\rho$) μεταξύ ζευγαριών των κατατάξεων που χρησιμοποιούμε ως υπόβαθρα αληθείας, για διαφορετικές τιμές του λόγου $\eta$.

| hep-th | $\eta = 1.2$ | | | $\eta = 1.4$ | | | $\eta = 1.6$ | | | $\eta = 1.8$ | | | $\eta = 2$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P-PR | I-CC | I-PR | P-PR | I-CC | I-PR | P-PR | I-CC | I-PR | P-PR | I-CC | I-PR | P-PR | I-CC | I-PR |
| P-CC | 0.690 | 0.649 | 0.376 | 0.747 | 0.747 | 0.453 | 0.775 | 0.806 | 0.520 | 0.790 | 0.846 | 0.560 | 0.794 | 0.861 | 0.596 |
| P-PR | | 0.381 | 0.568 | | 0.529 | 0.678 | | 0.609 | 0.743 | | 0.662 | 0.786 | | 0.689 | 0.840 |
| I-CC | | | 0.840 | | | 0.823 | | | 0.820 | | | 0.814 | | | 0.817 |

| APS | $\eta = 1.2$ | | | $\eta = 1.4$ | | | $\eta = 1.6$ | | | $\eta = 1.8$ | | | $\eta = 2$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P-PR | I-CC | I-PR | P-PR | I-CC | I-PR | P-PR | I-CC | I-PR | P-PR | I-CC | I-PR | P-PR | I-CC | I-PR |
| P-CC | 0.834 | 0.541 | 0.387 | 0.869 | 0.650 | 0.474 | 0.883 | 0.715 | 0.527 | 0.889 | 0.761 | 0.567 | 0.884 | 0.780 | 0.599 |
| P-PR | | 0.410 | 0.594 | | 0.547 | 0.665 | | 0.626 | 0.712 | | 0.680 | 0.747 | | 0.715 | 0.788 |
| I-CC | | | 0.904 | | | 0.895 | | | 0.887 | | | 0.880 | | | 0.876 |

| PMC | $\eta = 1.2$ | | | $\eta = 1.4$ | | | $\eta = 1.6$ | | | $\eta = 1.8$ | | | $\eta = 2$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P-PR | I-CC | I-PR | P-PR | I-CC | I-PR | P-PR | I-CC | I-PR | P-PR | I-CC | I-PR | P-PR | I-CC | I-PR |
| P-CC | 0.513 | 0.596 | 0.390 | 0.602 | 0.728 | 0.487 | 0.648 | 0.800 | 0.548 | 0.681 | 0.850 | 0.584 | 0.689 | 0.855 | 0.620 |
| P-PR | | 0.223 | 0.623 | | 0.392 | 0.769 | | 0.486 | 0.866 | | 0.562 | 0.840 | | 0.607 | 0.934 |
| I-CC | | | 0.871 | | | 0.852 | | | 0.844 | | | 0.812 | | | 0.822 |

| DBLP | $\eta = 1.2$ | | | $\eta = 1.4$ | | | $\eta = 1.6$ | | | $\eta = 1.8$ | | | $\eta = 2$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P-PR | I-CC | I-PR | P-PR | I-CC | I-PR | P-PR | I-CC | I-PR | P-PR | I-CC | I-PR | P-PR | I-CC | I-PR |
| P-CC | 0.757 | 0.662 | 0.517 | 0.803 | 0.757 | 0.591 | 0.825 | 0.810 | 0.635 | 0.837 | 0.844 | 0.672 | 0.835 | 0.854 | 0.693 |
| P-PR | | 0.436 | 0.761 | | 0.566 | 0.829 | | 0.641 | 0.866 | | 0.689 | 0.908 | | 0.723 | 0.924 |
| I-CC | | | 0.939 | | | 0.927 | | | 0.918 | | | 0.916 | | | 0.910 |

$k$. Στα πειράματα που εκτελούμε, θέτουμε το $k$ στις τιμές $\{5, 10, 50, 100, 500\}$, με το 50 να αποτελεί την τιμή βάσης.

**Συμβάσεις Ορολογίας.** Στο υπόλοιπο του κεφαλαίου, όσον αφορά τη διάτμηση των συνόλων δεδομένων, σύμφωνα με όσα αναπτύξαμε στην Ενότητα 3.2.1, θα αναφερόμαστε στο λόγο του μεγέθους (σε αριθμό δημοσιεύσεων) της μέλλουσας κατάστασης προς αυτόν της τρέχουσας κατάστασης απλά ως $\eta$. Επιπλέον, θα χρησιμοποιούμε ως συνώνυμο του όρου «μακροχρόνια απήχηση» τον όρο «επιρροή». Αντίστοιχα, θα χρησιμοποιούμε ως συνώνυμο του όρου «βραχυχρόνια απήχηση» τον όρο «δημοφιλία». Ακόμη, θα χρησιμοποιούμε απλά τον όρο «ακρίβεια», εννοώντας την ακρίβεια για ένα σύνολο $k$-σημαντικότερων αποτελεσμάτων. Αντίστοιχα θα χρησιμοποιούμε και τον όρο nDCG χωρίς να δηλώνουμε αναγκαστικά το σύνολο $k$. Τέλος, θα χρησιμοποιούμε τη μεταβλητή $k$ αναφερόμενοι στο σύνολο μεγέθους $|k|$ των σημαντικότερων δημοσιεύσεων βάσει θέσης κατάταξης.

## 4.2 Η Σχέση Μεταξύ Επιρροής - Δημοφιλίας

Το πρώτο ερευνητικό ερώτημα που θέσαμε αφορά τη διερεύνηση της σχέσης μεταξύ των δύο ειδών απήχησης. Για να διερευνήσουμε αυτή τη σχέση, υπολογίζουμε τις συσχετίσεις μεταξύ των κατατάξεων που παράγουν τα υπόβαθρα αληθείας που ορίζουν τη δημοφιλία και την επιρροή και τα οποία βασίζονται είτε στον αριθμό αναφορών, είτε στο PageRank (I-CC, I-PR, P-CC, ή P-PR). Στον Πίνακα 4.1 παρουσιάζουμε τη συσχέτιση μεταξύ των ζευγών των κατατάξεων, μετρημένη με το $\rho$ του Spearman, καθώς μεταβάλλουμε το $\eta$. Υπενθυμίζεται, πως ένας λόγος, π.χ., $\eta = 1.2$, σημαίνει ότι στη μέλλουσα κατάσταση περιλαμβάνονται 20% περισσότερες δημοσιεύσεις απ' ότι στην τρέχουσα κατάσταση. Η παρουσίαση μετρήσεων της συσχέτισης με χρήση του $\tau$ του Kendall παραλείπεται εδώ, καθώς οδήγησε σε παρόμοια αποτελέσματα.

Από τον Πίνακα 4.1 μπορούμε να αντλήσουμε μια σειρά από συμπεράσματα. Αρχικά, παρατηρούμε ότι όλα τα υπόβαθρα αληθείας είναι τουλάχιστον ασθενώς συσχετισμένα μεταξύ τους ($\rho > 0.2$), ενώ σε πολλές περιπτώσεις παρατηρείται ισχυρή συσχέτιση

($\rho > 0.8$).[7] Καθώς υπάρχει ένα εύρος τιμών συσχέτισης μεταξύ των υποβάθρων αληθείας, ειδικά μεταξύ αυτών που ορίζουν διαφορετικά είδη απήχησης, μπορούμε να συμπεράνουμε ότι, πράγματι, τα δύο αυτά υπόβαθρα αντικατοπτρίζουν διακριτές ιδιότητες (δημοφιλία - επιρροή).

Μια σημαντική παρατήρηση είναι ότι για διαφορετικά μέτρα κεντρικότητας που χρησιμοποιούνται για το ίδιο είδος απήχησης η μεταξύ τους συσχέτιση είναι ισχυρή. Συγκεκριμένα, παρατηρούμε $\rho > 0.6$ για την περίπτωση της δημοφιλίας και $\rho > 0.8$ για την περίπτωση της επιρροής. Η τελευταία παρατήρηση ενδεχομένως εξηγεί γιατί στη βιβλιογραφία δεν παρατηρείται πουθενά η χρήση του I-CC ως υπόβαθρου αληθείας, αλλά μόνο του I-PR. Από την άλλη, οι συσχετίσεις μεταξύ των δύο διαφορετικών ειδών απήχησης είναι ασθενέστερες. Όπως αναμένεται, οι συσχετίσεις μεταξύ των διαφορετικών ειδών απήχησης είναι ισχυρότερες όταν χρησιμοποιούν το ίδιο μέτρο κεντρικότητας, π.χ., 0.649 μεταξύ P-CC, I-CC έναντι 0.376 μεταξύ P-CC και I-PR στο σύνολο δεδομένων hep-th και $\eta = 1.2$.

Μια επιπλέον τάση που διακρίνουμε είναι η αυξανόμενη συσχέτιση μεταξύ δημοφιλίας και επιρροής καθώς αυξάνει η τιμή του λόγου $\eta$. Με την αύξηση του $\eta$ η μέλλουσα κατάσταση, που περιγράφεται από τον πίνακα $\boldsymbol{A}(t_c + T)$, αυξάνει σε σχέση με την τρέχουσα κατάσταση, που περιγράφεται από τον πίνακα $\boldsymbol{A}(t_c)$ και ο οποίος είναι σταθερός. Επομένως, οι αναφορές και οι αλυσίδες αναφορών στον πίνακα $\boldsymbol{A}(t_c+T) - \boldsymbol{A}(t_c)$ τείνουν να προσεγγίσουν περισσότερο αυτές του $\boldsymbol{A}(t_c + T)$, με αποτέλεσμα τα μέτρα κεντρικότητας που υπολογίζονται στον πρώτο γράφο να προσεγγίζουν περισσότερο τα μέτρα κεντρικότητας που υπολογίζονται στον δεύτερο.

Συνολικά συμπεραίνουμε ότι η δημοφιλία και η επιρροή αντικατοπτρίζουν διακριτές ιδιότητες, οι οποίες ωστόσο είναι σε κάποιο βαθμό συσχετισμένες.[8] Τα δύο διαφορετικά μέτρα κεντρικότητας, ο αριθμός αναφορών και το PageRank, παράγουν σχετικά διακριτές κατατάξεις όσον αφορά την δημοφιλία, ειδικά για μικρές τιμές του $\eta$ - μια ρύθμιση που είναι προτιμότερη για τη μέτρηση της δημοφιλίας. Από την άλλη, όσον αφορά την επιρροή, τα δύο μέτρα κεντρικότητας παράγουν πολύ παρόμοιες κατατάξεις, ειδικά για μεγάλες τιμές του $\eta$ - μια ρύθμιση που είναι προτιμότερη για τη μέτρηση της συνολικής επιρροής.

## 4.3 Αξιολόγηση Αποτελεσματικότητας Κατάταξης

### 4.3.1 Αξιολόγηση Βάσει Επιρροής

Σε αυτή την ενότητα μελετάμε το δεύτερο ερευνητικό ερώτημα που τέθηκε στην Ενότητα 4.1.1, με αναφορά στην παραγωγή κατατάξεων με βάση την επιρροή. Η αξιολόγηση αρχικά πραγματοποιείται λαμβάνοντας υπόψη και τα δύο μέτρα κεντρικότητας που την ορίζουν. Σε ακόλουθες υποενότητες εστιάζουμε στην επιρροή, όπως ορίζεται από το I-PR, καθώς εμφανίζεται συχνότερα στη βιβλιογραφία. Ένας επιπλέον λόγος είναι ότι διαισθητικά, η επιρροή μετρημένη μέσω του I-PR φαίνεται μια πιο λογική επιλογή, σε σύγκριση με το I-CC, διότι το PageRank λαμβάνει υπόψη όχι μόνο την άμεση επιρροή

[7] Χρησιμοποιούμε τις ερμηνείες για το πόσο ισχυρή είναι μια συσχέτιση με βάση όσα δίνει ο Evans [20].

[8] Κάποιο επίπεδο συσχέτισης είναι αναμενόμενο, καθώς, για παράδειγμά, πολλές δημοσιεύσεις που έχουν μεγάλη επιρροή παραμένουν και δημοφιλείς.

Πίνακας 4.2: hep-th: μετρικές για τα I-CC, I-PR· $\eta = 1.6$, $k = 50$.

| **hep-th** | **I-CC** | | | | **I-PR** | | | |
|---|---|---|---|---|---|---|---|---|
| | $\rho$ | $\tau$ | **Ακρίβεια** | **nDCG** | $\rho$ | $\tau$ | **Ακρίβεια** | **nDCG** |
| PR | 0.734 | 0.571 | 0.480 | 0.645 | **0.892** | **0.768** | 0.780 | 0.909 |
| NPR | 0.675 | 0.513 | 0.260 | 0.482 | 0.874 | 0.726 | 0.480 | 0.793 |
| CR | 0.752 | 0.571 | 0.720 | 0.880 | 0.822 | 0.652 | **0.880** | **0.967** |
| FR | 0.512 | 0.380 | **0.740** | 0.865 | 0.419 | 0.300 | 0.780 | 0.958 |
| ECM | 0.830 | 0.679 | 0.400 | 0.795 | 0.684 | 0.509 | 0.320 | 0.665 |
| RAM | **0.836** | **0.689** | 0.700 | **0.946** | 0.698 | 0.524 | 0.480 | 0.802 |
| NR | 0.307 | 0.216 | 0.200 | 0.470 | 0.338 | 0.242 | 0.300 | 0.622 |

Πίνακας 4.3: APS: μετρικές για τα I-CC, I-PR· $\eta = 1.6$, $k = 50$.

| **APS** | **I-CC** | | | | **I-PR** | | | |
|---|---|---|---|---|---|---|---|---|
| | $\rho$ | $\tau$ | **Ακρίβεια** | **nDCG** | $\rho$ | $\tau$ | **Ακρίβεια** | **nDCG** |
| PR | **0.760** | **0.603** | 0.300 | 0.611 | **0.897** | **0.781** | 0.740 | 0.955 |
| MPR | 0.734 | 0.576 | 0.340 | 0.626 | 0.888 | 0.752 | 0.760 | 0.952 |
| CR | 0.573 | 0.437 | 0.600 | 0.827 | 0.486 | 0.357 | **0.780** | **0.958** |
| FR | 0.486 | 0.361 | **0.640** | 0.840 | 0.377 | 0.269 | 0.620 | 0.847 |
| ECM | 0.692 | 0.534 | 0.580 | **0.846** | 0.577 | 0.419 | 0.340 | 0.664 |
| RAM | 0.692 | 0.534 | 0.600 | 0.837 | 0.576 | 0.419 | 0.340 | 0.663 |
| NR | 0.169 | 0.120 | 0.200 | 0.290 | 0.030 | 0.022 | 0.220 | 0.379 |

μιας δημοσίευσης στην άλλη, αλλά και την έμμεση επιρροή που ασκούν μεταξύ τους οι δημοσιεύσεις μέσω αλυσίδων αναφορών. Στις ακόλουθες υποενότητες εξετάζουμε με τη σειρά:

- Συνολικά την αποτελεσματικότητα κάθε μεθόδου στην κατάταξη δημοσιεύσεων με βάση την επιρροή.
- Πώς μεταβάλλεται η αποτελεσματικότητα των μεθόδων κατάταξης στη παραγωγή κατατάξεων βάσει επιρροής, καθώς μεταβάλλεται ο λόγος $\eta$, βάσει της συνολικής συσχέτισης και του nDCG.
- Πώς μεταβάλλεται η αποτελεσματικότητα των μεθόδων κατάταξης βάσει του nDCG, για διαφορετικές τιμές του $k$.

#### 4.3.1.1 Επισκόπηση Αποτελεσματικότητας Μεθόδων Κατάταξης Βάσει Επιρροής

Στο πείραμα αυτής της υποενότητας θέτουμε τον λόγο $\eta$ στη τιμή βάσης του ($\eta = 1.6$) και υπολογίζουμε όλες τις μετρικές ($\rho$, $\tau$, ακρίβεια και nDCG για $k = 50$) που ποσοτικοποιούν την αποτελεσματικότητα όλων των μεθόδων κατάταξης ως προς την επίτευξη της παραγωγής μιας κατάταξης βάσει της επιρροής, είτε αυτή μετριέται με τον αριθμό αναφορών (I-CC), είτε με το PageRank (I-PR). Οι Πίνακες 4.2–4.5 παρουσιάζουν τα αποτελέσματα για κάθε σύνολο δεδομένων. Παρατηρούμε ότι, όσον αφορά το I-CC, η μέθοδος RAM έχει τα καλύτερα αποτελέσματα από όλες τις μεθόδους στα περισσότερα σύνολα δεδομένων και για τις περισσότερες μετρικές. Όσον αφορά το I-PR, οι μέθοδοι PR και CR υπερισχύουν, με την πρώτη να πετυχαίνει την καλύτερη συνολική συσχέτιση (μετρικές $\rho$ και $\tau$) και τη δεύτερη να πετυχαίνει τα καλύτερα αποτελέσματα όσο αφορά την κατάταξη των πρώτων 50 αποτελεσμάτων (ακρίβεια και nDCG).

Πίνακας 4.4: PMC: μετρικές για τα I-CC, I-PR· $\eta = 1.6$, $k = 50$.

| **PMC** | **I-CC** | | | | **I-PR** | | | |
|---|---|---|---|---|---|---|---|---|
| | $\rho$ | $\tau$ | **Ακρίβεια** | **nDCG** | $\rho$ | $\tau$ | **Ακρίβεια** | **nDCG** |
| PR | 0.726 | 0.591 | 0.360 | 0.652 | **0.818** | **0.694** | 0.840 | 0.969 |
| NPR | 0.708 | 0.570 | 0.360 | 0.649 | 0.814 | 0.682 | 0.780 | 0.952 |
| CR | 0.563 | 0.426 | 0.580 | 0.842 | 0.603 | 0.457 | **0.900** | **0.990** |
| FR | 0.261 | 0.196 | 0.580 | 0.803 | 0.219 | 0.161 | 0.820 | 0.977 |
| ECM | **0.787** | 0.677 | 0.800 | 0.967 | 0.751 | 0.594 | 0.440 | 0.797 |
| RAM | **0.787** | **0.679** | **0.820** | **0.969** | 0.751 | 0.596 | 0.420 | 0.794 |
| NR | 0.183 | 0.134 | 0.360 | 0.555 | 0.226 | 0.160 | 0.580 | 0.812 |
| YR | 0.614 | 0.469 | 0.400 | 0.693 | 0.618 | 0.467 | 0.760 | 0.938 |
| WSDM | 0.567 | 0.432 | 0.160 | 0.478 | 0.465 | 0.326 | 0.140 | 0.437 |
| WC | 0.772 | 0.649 | 0.660 | 0.895 | 0.737 | 0.574 | 0.380 | 0.715 |

Πίνακας 4.5: DBLP: μετρικές για τα I-CC, I-PR· $\eta = 1.6$, $k = 50$.

| **DBLP** | **I-CC** | | | | **I-PR** | | | |
|---|---|---|---|---|---|---|---|---|
| | $\rho$ | $\tau$ | **Ακρίβεια** | **nDCG** | $\rho$ | $\tau$ | **Ακρίβεια** | **nDCG** |
| PR | 0.811 | 0.673 | 0.480 | 0.717 | **0.884** | **0.778** | 0.820 | 0.981 |
| NPR | 0.797 | 0.655 | 0.440 | 0.726 | 0.880 | 0.763 | 0.740 | 0.965 |
| CR | 0.549 | 0.413 | 0.740 | 0.938 | 0.537 | 0.402 | **0.860** | **0.988** |
| FR | 0.389 | 0.294 | 0.780 | 0.947 | 0.349 | 0.257 | 0.720 | 0.950 |
| ECM | **0.845** | 0.726 | **0.820** | **0.966** | 0.812 | 0.656 | 0.520 | 0.800 |
| RAM | **0.845** | **0.727** | **0.820** | **0.966** | 0.812 | 0.656 | 0.520 | 0.800 |
| NR | 0.101 | 0.074 | 0.400 | 0.710 | 0.050 | 0.037 | 0.440 | 0.714 |
| YR | 0.627 | 0.490 | 0.620 | 0.836 | 0.682 | 0.553 | 0.800 | 0.970 |
| WSDM | 0.616 | 0.465 | 0.580 | 0.698 | 0.593 | 0.437 | 0.440 | 0.688 |
| WC | 0.839 | 0.714 | 0.480 | 0.630 | 0.833 | 0.682 | 0.320 | 0.499 |

Ας εξετάσουμε αναλυτικότερα τα αποτελέσματα όταν χρησιμοποιούμε ως υπόβαθρο αληθείας το I-CC. Αναμένουμε οι μέθοδοι που βασίζονται στη μέτρηση του αριθμού αναφορών να είναι αποτελεσματικές όσο αφορά τη συσχέτιση με το I-CC (καθώς χρησιμοποιούν το ίδιο μέτρο κεντρικότητας), όπως πράγματι παρατηρούμε ότι συμβαίνει με τις μεθόδους RAM και WC. Αν εστιάσουμε, ωστόσο, στην ακρίβεια και το nDCG ($k = 50$) παρατηρούμε ότι η μέθοδος RAM υπερισχύει της WC. Καθώς και οι δύο μέθοδοι χρησιμοποιούν χρονικά ενήμερα βάρη στον πίνακα γειτνίασης, μπορούμε να συμπεράνουμε ότι η χρήση της ηλικίας αναφοράς (RAM) είναι πιο αποτελεσματική από την χρήση του χρονικού χάσματος αναφοράς. Επιπλέον μπορούμε να κάνουμε κάποιες παρατηρήσεις σχετικά με τις μεθόδους που βασίζονται στο PageRank. Αν και οι PR και CR δεν πετυχαίνουν καλή συνολική συσχέτιση, ωστόσο είναι ιδιαίτερα αποτελεσματικές, όσον αφορά την ακρίβεια και το nDCG. Υποθέτουμε ότι αυτό δεν συμβαίνει εξαιτίας της χρήσης του PageRank (καθώς άλλες μέθοδοι βασισμένες στο PageRank δεν είναι εξίσου αποτελεσματικές), αλλά στη χρήση πιθανοτήτων επιλογής που φθίνουν με την ηλικία των δημοσιεύσεων. Ωστόσο, ο συνδυασμός τέτοιων πιθανοτήτων επιλογής μαζί με ένα χρονικά ενήμερο πίνακα γειτνίασης δεν οδηγεί σε καλύτερα αποτελέσματα, όπως φαίνεται από την περίπτωση της μεθόδου NR. Επιπλέον, η μέθοδος ECM, παρότι λαμβάνει υπόψη μονοπάτια αναφορών, είναι περίπου εξίσου αποτελεσματική με τη μέθοδο RAM. Αποδίδουμε αυτή τη συμπεριφορά σε δύο παράγοντες: αφενός, τόσο η RAM όσο και η ECM χρησιμοποιούν τις ίδιες χρονικές ποσότητες στον πίνακα γειτνίασης

και αφετέρου, η μέθοδος ECM είναι ρυθμισμένη[9] έτσι ώστε το βάρος μεγαλύτερων μονοπατιών αναφορών προς μια δημοσίευση, στον υπολογισμό των βαθμολογιών, να φθίνει γρήγορα ανάλογα με το μέγεθος του μονοπατιού. Έτσι η μέθοδος ECM σε μεγάλο βαθμό «εκφυλίζεται» στη μέθοδο RAM, καθώς οι βαθμολογίες που υπολογίζει εξαρτιόνται κυρίως από τις άμεσες αναφορές. Παρατηρούμε αυτή τη στενή σχέση μεταξύ των RAM και ECM σε όλα τα σύνολα δεδομένων, για όλα τα είδη απήχησης και όλες τις μετρικές αξιολόγησης.

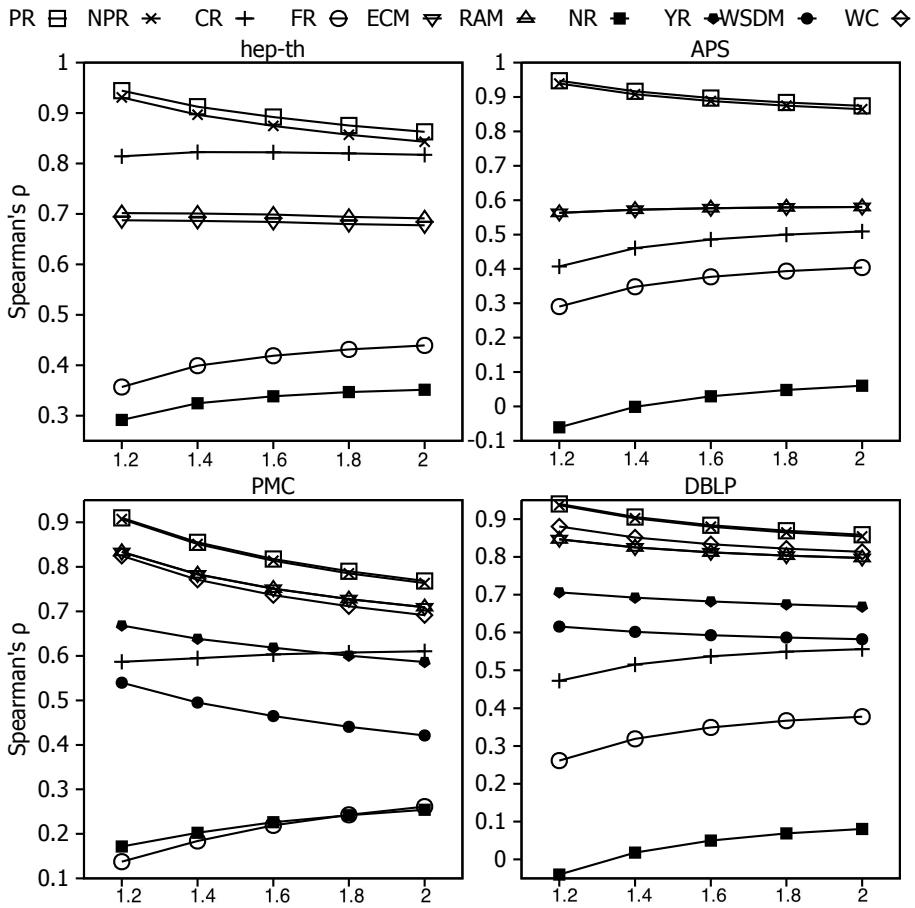
Εξετάζουμε στη συνέχεια την περίπτωση που ως υπόβαθρο αληθείας χρησιμοποιούμε το I-PR, σενάριο στο οποίο αναμένουμε να υπερισχύουν οι μέθοδοι που βασίζονται στο PageRank. Η διαίσθηση αυτή επιβεβαιώνεται από τα δεδομένα, καθώς παρατηρούμε ότι οι μέθοδοι PR και NPR είναι οι αποτελεσματικότερες όσον αφορά τη συνολική κατάταξη, ενώ οι CR και FR υπερισχύουν με βάση την ακρίβεια και το nDCG. Το απλό PageRank πετυχαίνει τα βέλτιστα αποτελέσματα όσον αφορά τη συσχέτιση που μετράμε με τις μετρικές $\rho$ και $\tau$, επειδή η γενική δομή του δικτύου αναφορών που περιγράφεται από τον πίνακα $\boldsymbol{A}(t_c)$, στον οποίο εφαρμόζεται η μέθοδος, είναι σε μεγάλο βαθμό όμοια με αυτή του δικτύου αναφορών που περιγράφεται από τον πίνακα $\boldsymbol{A}(t_c+T)$, στον οποίο υπολογίζεται το I-PR. Αν και γενικά το PageRank είναι αποτελεσματικό, δεν είναι η καλύτερη μέθοδος όσον αφορά την ακρίβεια και το nDCG. Βάσει των μετρικών αυτών είναι πιο αποτελεσματική η μέθοδος CR εξαιτίας ων εκθετικά φθινουσών με την ηλικία των δημοσιεύσεων πιθανοτήτων επιλογής που χρησιμοποιεί. Τέτοιες πιθανότητες επιλογής αναιρούν τη μεροληψία που έχει το απλό PR υπέρ των παλαιότερων δημοσιεύσεων. Οι παρατηρήσεις για τη μέθοδο CR ισχύουν και για τη μέθοδο FR που χρησιμοποιεί παρόμοιες πιθανότητες επιλογής.

#### 4.3.1.2 Μεταβάλλοντας το λόγο $\eta$

Σε αυτό το πείραμα μεταβάλλουμε το λόγο $\eta$ και μετράμε την αποτελεσματικότητα όλων των μεθόδων σε σχέση με το υπόβαθρο αληθείας ορισμένο ως I-PR. Στο Σχήμα 4.1 παρουσιάζουμε τις τιμές $\rho$ του Spearman για κάθε μέθοδο και για κάθε σύνολο δεδομένων. Καθώς παρατηρήσαμε παρόμοια αποτελέσματα για το $\tau$ του Kendall, αυτά παραλείπονται. Όπως παρατηρήσαμε και προηγουμένως, οι μέθοδοι που βασίζονται στο PageRank και ιδιαίτερα οι PR και NPR είναι οι αποτελεσματικότερες, όσο αφορά την επίτευξη καλής συσχέτισης με το υπόβαθρο αληθείας. Υπενθυμίζουμε ότι η μέθοδος NPR είναι μια βασική παραλλαγή του PR που δεν εισάγει χρονικούς παράγοντες και δε χρησιμοποιεί επιπλέον πληροφορία όπως μεταδεδομένα. Όσο μικρότερος είναι ο λόγος $\eta$, τόσο μεγαλύτερη η συσχέτιση που πετυχαίνουν οι μέθοδοι με το υπόβαθρο αληθείας. Αυτό συμβαίνει, διότι η διαφορά μεταξύ του τρέχοντος δικτύου και του μελλοντικού δικτύου αναφορών, στο οποίο υπολογίζεται η κατάταξη με βάση το I-PR, είναι μικρότερη για μικρότερες τιμές του $\eta$.

Παρατηρούμε επιπλέον μια ομάδα μεθόδων, τις CR, FR, και NR, οι οποίες φαίνεται να ευνοούνται καθώς το $\eta$ αυξάνει. Αυτές οι μέθοδοι, που βασίζονται στο PageRank, κάνουν χρήση πιθανοτήτων επιλογής που φθίνουν με την ηλικία των δημοσιεύσεων και συνολικά δεν είναι πολύ αποτελεσματικές, όσον αφορά τη συνολική συσχέτιση (με εξαίρεση τη μέθοδο CR). Αυτό συμβαίνει για τον εξής λόγο: καθώς αυξάνει το $\eta$ η κατάταξη του υπόβαθρου αληθείας αποκλίνει όλο και περισσότερο από την απλή κατάταξη με βάση το PageRank στο δίκτυο αναφορών που αντιστοιχεί στην τρέχουσα κατάσταση. Αυτό συμβαίνει γιατί οι νεότερες δημοσιεύσεις αρχίζουν να συσσωρεύουν

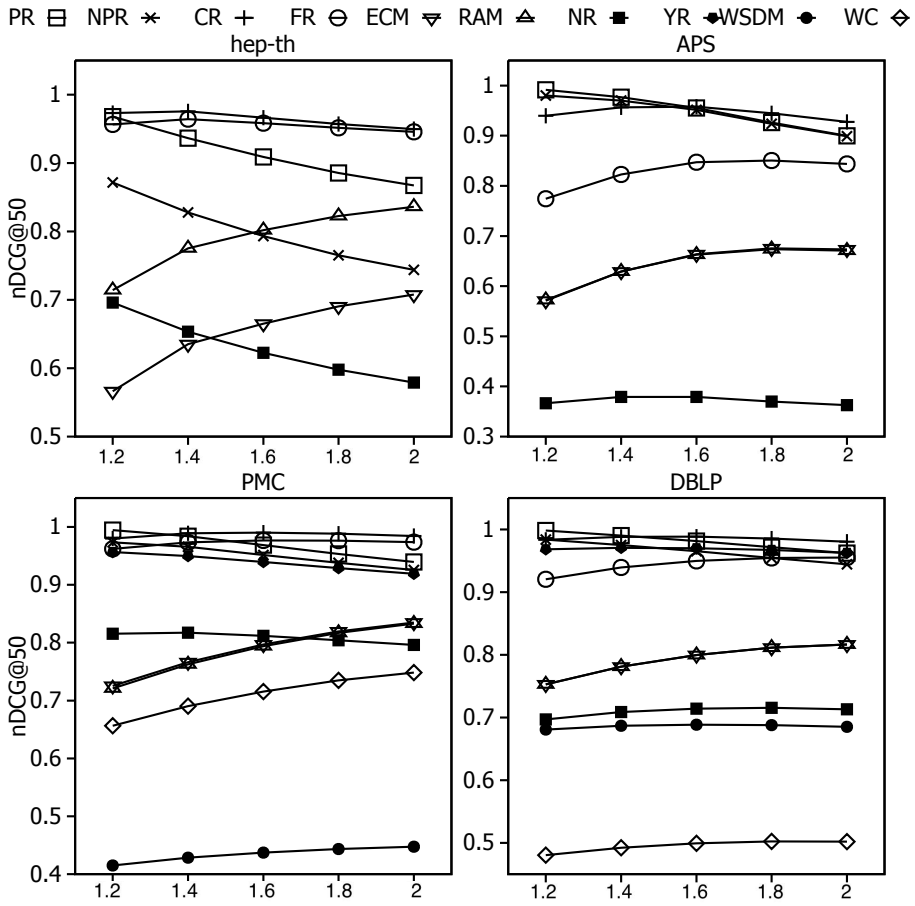
[9]Βάσει των ρυθμίσεων που παρουσιάζονται από τους συγγραφείς τις.




Σχήμα 4.1: Συσχέτιση της κατάταξης κάθε μεθόδου με αυτήν του I-PR, καθώς μεταβάλλουμε το $\eta$.

αναφορές και βελτιώνεται η σχετική τους επιρροή σε σύγκριση με αυτή των παλαιότερων δημοσιεύσεων. Το φαινόμενο αυτό «προβλέπεται» άμεσα από τις πιθανότητες επιλογής που χρησιμοποιούν αυτές οι μέθοδοι, οι οποίες προωθούν *όλες* τις νεότερες δημοσιεύσεις. Ωστόσο, δεν έχουν όλες οι πρόσφατες δημοσιεύσεις την ίδια επιρροή και έτσι εξηγούνται οι σχετικά χαμηλές τιμές στις συσχετίσεις. Η αύξηση της τιμής του $\eta$ συνεπάγεται ότι εκείνες οι λίγες πρόσφατες δημοσιεύσεις που πράγματι έχουν επιρροή αρχίζουν και αντανακλούν την απήχησή τους στη κατάταξη και έτσι η αποτελεσματικότητα των μεθόδων που ρητά τις προωθούν (μαζί με άλλες νεότερες δημοσιεύσεις που δεν έχουν αντίστοιχη επιρροή) αυξάνει.

Στο Σχήμα 4.2 παρουσιάζουμε την τιμή nDCG που πετυχαίνει κάθε μέθοδος ($k = 50$) με βάση το υπόβαθρο αληθείας, καθώς μεταβάλλεται ο λόγος $\eta$. Παρόμοια αποτελέσματα δίνουν και οι μετρήσεις της ακρίβειας. Το απλό PR βρίσκεται ανάμεσα στις αποτελεσματικότερες μεθόδους, ωστόσο η αποτελεσματικότητά του πέφτει καθώς αυξάνει ο λόγος $\eta$. Λαμβάνοντας υπόψη όλες τις τιμές του $\eta$, το CR φαίνεται να είναι συνολικά η αποτελεσματικότερη μέθοδος, συχνά ακολουθούμενη από τη μέθοδο FR. Όπως αναφέρθηκε και νωρίτερα, οι μέθοδοι CR, FR και YR χρησιμοποιούν πιθανότητες επιλογής που φθίνουν με την ηλικία των δημοσιεύσεων, για να προωθήσουν νεότερες δημοσιεύσεις σε καλύτερες θέσεις κατάταξης. Αναμένουμε, ότι μια δημοσίευση θα βρεθεί στις καλύτερες θέσεις κατάταξης με βάση την επιρροή εφόσον είχε ήδη μεγάλη επιρροή (στο τρέχον δίκτυο αναφορών), ή εάν είχε κάποια μέτρια επιρροή και συσσωρεύει πολλές αναφορές σε ένα σε ενα πρόσφατο κοντινό χρονικό διάστημα. Οι

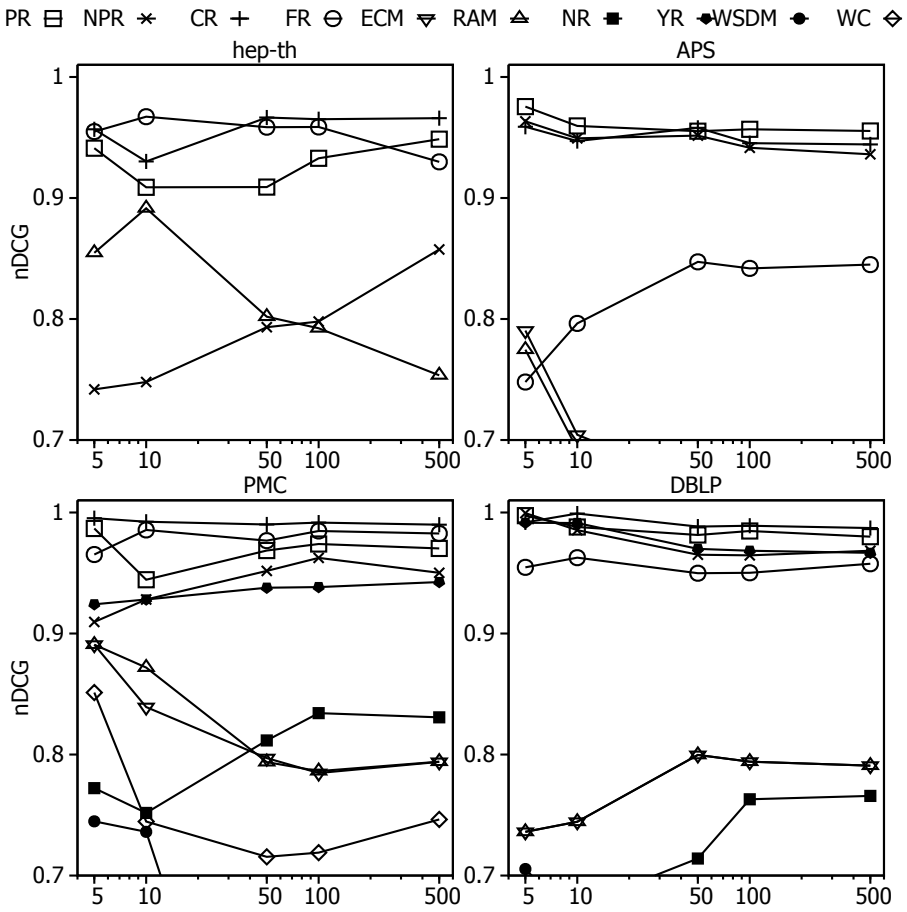



Σχήμα 4.2: nDCG@50 για τη κατάταξη κάθε μεθόδου, με αναφορά στην κατάταξη του I-PR, μεταβάλλοντας την τιμή του $\eta$.

μέθοδοι CR και FR μπορούν να εντοπίσουν και τις δύο αυτές περιπτώσεις, ενώ το PR μπορεί να εντοπίσει μόνο την πρώτη. Μια επιπλέον παρατήρηση είναι ότι η χρήση χρονικά ενήμερων πινάκων γειτνίασης/μεταβάσεων, που γίνεται από τις μεθόδους RAM, ECM, WC και NR, δεν φαίνεται να ευνοεί στον εντοπισμό των δημοσιεύσεων με τη μεγαλύτερη επιρροή, όπως αυτή ορίζεται με το I-PR. Η αποτελεσματικότητα τους, ωστόσο, αυξάνει γρήγορα καθώς αυξάνει το $\eta$, πράγμα που υποδηλώνει ότι ένας χρονικά ενήμερος μηχανισμός είναι σημαντικός. Ωστόσο, η αύξηση στην αποτελεσματικότητα που προσφέρει φαίνεται να έχει κάποιο άνω όριο. Συνολικά φαίνεται ότι οι πιθανότητες επιλογής που φθίνουν με την ηλικία των δημοσιεύσεων είναι ο πιο αποτελεσματικός μηχανισμός στο σενάριο που εξετάζουμε.

#### 4.3.1.3 Μεταβάλλοντας το $k$

Στο τελευταίο πείραμα που αφορά την κατάταξη με βάση την επιρροή, μετράμε την τιμή του nDCG που πετυχαίνει κάθε μέθοδος, μεταβάλλοντας το $k$ στις τιμές $\{5, 10, 50, 100, 500\}$, διατηρώντας το $\eta$ σταθερό στην τιμή αναφοράς $\eta = 1.6$. Στο Σχήμα 4.3 παρουσιάζονται τα αποτελέσματα. Συνολικά βλέπουμε ότι οι μέθοδοι που ήταν αποτελεσματικές στα προηγούμενα πειράματα (PR, NPR, CR, FR, YR) είναι σχετικά σταθερές καθώς μεταβάλλεται το $k$, ενώ η αποτελεσματικότητα των υπολοίπων μεθόδων παρουσιάζει μεγάλες διακυμάνσεις.




Σχήμα 4.3: nDCG της κατάταξης κάθε μεθόδου, με αναφορά στο υπόβαθρο αληθείας I-PR, υπολογισμένο για διάφορες τιμές των $k$ σημαντικότερων αποτελεσμάτων· $\eta = 1.6$.

### 4.3.2 Αξιολόγηση Βάσει Δημοφιλίας

Σε αυτή την ενότητα συνεχίζουμε τη διερεύνηση του δεύτερου ερευνητικού ερωτήματος, αυτή τη φορά εστιάζοντας στην κατάταξη με βάση τη δημοφιλία. Συγκρίνουμε την αποτελεσματικότητα των μεθόδων ως προς την παραγωγή κατατάξεων με βάση τη δημοφιλία, όπως αυτή ορίζεται από τον αριθμό αναφορών και το PageRank. Στις επόμενες ενότητες πραγματοποιούμε με τη σειρά πειράματα για να διερευνήσουμε:

- Συνολικά την αποτελεσματικότητα κάθε μεθόδου στη κατάταξη δημοσιεύσεων με βάση τη δημοφιλία.
- Πως μεταβάλλεται η αποτελεσματικότητα των μεθόδων κατάταξης στη παραγωγή κατατάξεων βάσει της δημοφιλίας, καθώς μεταβάλλεται ο λόγος $\eta$, βάσει της συνολικής συσχέτισης και του nDCG.
- Πως μεταβάλλεται η αποτελεσματικότητα των μεθόδων κατάταξης βάσει του nDCG, για διαφορετικές τιμές του $k$.

Στα πειράματα αυτής της ενότητας εστιάζουμε περισσότερο στην κατάταξη με βάση τη δημοφιλία, όπως αυτή ορίζεται από το P-CC, καθώς αυτό χρησιμοποιείται συχνότερα στη βιβλιογραφία. Επιπλέον, θεωρούμε ότι διαισθητικά είναι πιο κατάλληλο μέτρο για τη δημοφιλία, καθώς αντανακλά την άμεση επίδραση που έχει μια δημοσίευση πάνω σε άλλες, σε αντίθεση με την περίπτωση που χρησιμοποιούμε το PageRank, όπου λαμβάνεται υπόψη και η έμμεση επίδραση μέσω αλυσίδων αναφορών.

#### 4.3.2.1 Επισκόπηση Αποτελεσματικότητας Μεθόδων Κατάταξης Βάσει Δημοφιλίας

Στο πρώτο πείραμα, θέτουμε το λόγο $\eta$ στη τιμή βάσης ($\eta = 1.6$) και υπολογίζουμε όλες τις μετρικές αξιολόγησης ($\rho$, $\tau$, ακρίβεια και nDCG για $k = 50$) που μετρούν την αποτελεσματικότητα των μεθόδων κατάταξης με αναφορά στο υπόβαθρο αληθείας για τη δημοφιλία, όπως ορίζεται μέσω του αριθμού αναφορών(P-CC) και του PageRank (P-PR). Οι Πίνακες 4.6–4.9 παρουσιάζουν τα αποτελέσματα για κάθε σύνολο δεδομένων. Γενικά, παρατηρούμε ότι για τη δημοφιλία ορισμένη μέσω του P-CC, οι μέθοδοι RAM και ECM είναι οι αποτελεσματικότερες για όλες τις μετρικές αξιολόγησης, στη πλειοψηφία των συνόλων δεδομένων. Στην περίπτωση που ορίζουμε την κατάταξη της δημοφιλίας βάσει του P-PR, η μέθοδος CR πετυχαίνει την καλύτερη συνολική συσχέτιση, ενώ οι μέθοδοι RAM, ECM και FR είναι αποτελεσματικότερες όσον αφορά την ακρίβεια και το nDCG.

Εξετάζοντας πιο λεπτομερώς την περίπτωση που ορίζουμε τη δημοφιλία μέσω του P-CC, μπορούμε να πούμε ότι η αποτελεσματικότητα των μεθόδων RAM και ECM οφείλεται στο ότι κυρίως κατατάσσουν τις δημοσιεύσεις με βάση τις αναφορές που έλαβαν πρόσφατα, κάτι το οποίο φαίνεται να είναι καλή ένδειξη του κατά πόσο θα εξακολουθήσουν να λαμβάνουν αναφορές στο προσεχές μέλλον. Παρατηρούμε, ωστόσο, μια εξαίρεση στην περίπτωση του συνόλου δεδομένων APS, όπου η μέθοδος CR υπερισχύει των RAM, ECM όσον αφορά τη συσχέτιση με το υπόβαθρο αληθείας. Αυτό το φαινόμενο οφείλεται στη φύση του συνόλου δεδομένων APS, όπου η συσχέτιση των κατατάξεων με βάση τα P-CC, P-PR είναι αρκετά υψηλή (βλ. Πίνακα 4.1), ενώ, όπως θα συζητήσουμε στη συνέχεια, η μέθοδος CR πετυχαίνει υψηλή συσχέτιση με την κατάταξη του I-PR. Συνολικά, σημειώνουμε ότι η χρήση χρονικών παραγόντων είναι σημαντική στο σενάριο της κατάταξης με βάση τη δημοφιλία. Επίσης, σε αντίθεση με το σενάριο της επιρροής, στο σενάριο της δημοφιλίας οι μέθοδοι που βασίζονται στο PageRank και χρησιμοποιούν πιθανότητες επιλογής με χρονικούς παράγοντες (ιδιαίτερα οι CR, FR) πετυχαίνουν καλή αποτελεσματικότητα.

Όσον αφορά την περίπτωση της δημοφιλίας με βάση το P-PR, παρατηρούμε ότι η μέθοδος CR πετυχαίνει την καλύτερη συνολική συσχέτιση. Αυτό συμβαίνει γιατί ο «σκόπιμος ερευνητής» που προσομοιώνει, έχει συμπεριφορά παρόμοια με αυτή του τυχαίου ερευνητή που ορίζεται από το P-PR. Συγκεκριμένα, η μέθοδος CR κάνει την παραδοχή ότι οι ερευνητές προτιμούν να διαβάζουν πρόσφατες δημοσιεύσεις, ενώ το P-PR προσομοιώνει έναν ερευνητή κυρίως φτάνει στις ίδιες δημοσιεύσεις, μέσω των νεότερων που τις αναφέρουν. Καθώς οι πρόσφατες δημοσιεύσεις εκδίδονται σε κοντινό χρονικό διάστημα με αυτές του προσεχούς μέλλοντος, οι τάση συσσώρευσης αναφορών στη μια και την άλλη ομάδα δημοσιεύσεων είναι παρόμοια. Επομένως οι δύο διαδικασίες παράγουν παρόμοιες κατατάξεις. Σημειώνουμε, ωστόσο, ότι οι συσχετίσεις που παρατηρούμε με τα υπόβαθρα αληθείας, είναι σημαντικά ασθενέστερες σε σχέση με την περίπτωση της κατάταξης με βάση την επιρροή. Όσον αφορά την ακρίβεια και το nDCG, παρατηρούμε ότι οι μέθοδοι που χρησιμοποιούν ως χρονική ποσότητα την ηλικία αναφορών (οι RAM, ECM) είναι οι αποτελεσματικότερες. Έπονται, λίγο λιγότερο αποτελεσματικές οι μέθοδοι που χρησιμοποιούν πιθανότητες επιλογής με βάση την ηλικία των δημοσιεύσεων (CR, FR). Κατά παρόμοιο τρόπο με τις χρονικά εξαρτημένες πιθανότητες επιλογής, η χρήση βαρών με βάση την ηλικία αναφοράς ευνοεί τις πιο πρόσφατες δημοσιεύσεις.

Συνολικά, παρατηρούμε ότι για την επίτευξη καλής κατάταξης με βάση τη δημοφιλία, απαιτείται η μετρίαση της μεροληψίας υπέρ των παλαιότερων δημοσιεύσεων. Φαίνεται

Πίνακας 4.6: hep-th: μετρικές για τα P-CC, P-PR· $\eta = 1.6$, $k = 50$.

| **hep-th** | **P-CC** | | | | **P-PR** | | | |
|---|---|---|---|---|---|---|---|---|
| | $\rho$ | $\tau$ | **Ακρίβεια** | **nDCG** | $\rho$ | $\tau$ | **Ακρίβεια** | **nDCG** |
| PR | 0.301 | 0.217 | 0.300 | 0.332 | 0.243 | 0.166 | 0.240 | 0.353 |
| NPR | 0.249 | 0.179 | 0.200 | 0.240 | 0.227 | 0.161 | 0.180 | 0.410 |
| CR | 0.561 | 0.416 | 0.500 | 0.638 | **0.466** | **0.323** | 0.400 | 0.566 |
| FR | 0.548 | 0.407 | 0.540 | 0.659 | 0.453 | 0.313 | **0.480** | 0.620 |
| ECM | 0.578 | 0.437 | 0.400 | 0.776 | 0.319 | 0.219 | 0.360 | 0.531 |
| RAM | **0.601** | **0.460** | **0.580** | **0.855** | 0.360 | 0.251 | 0.440 | **0.640** |
| NR | 0.311 | 0.223 | 0.200 | 0.370 | 0.339 | 0.231 | 0.220 | 0.494 |

Πίνακας 4.7: APS: μετρικές για τα P-CC, P-PR· $\eta = 1.6$, $k = 50$.

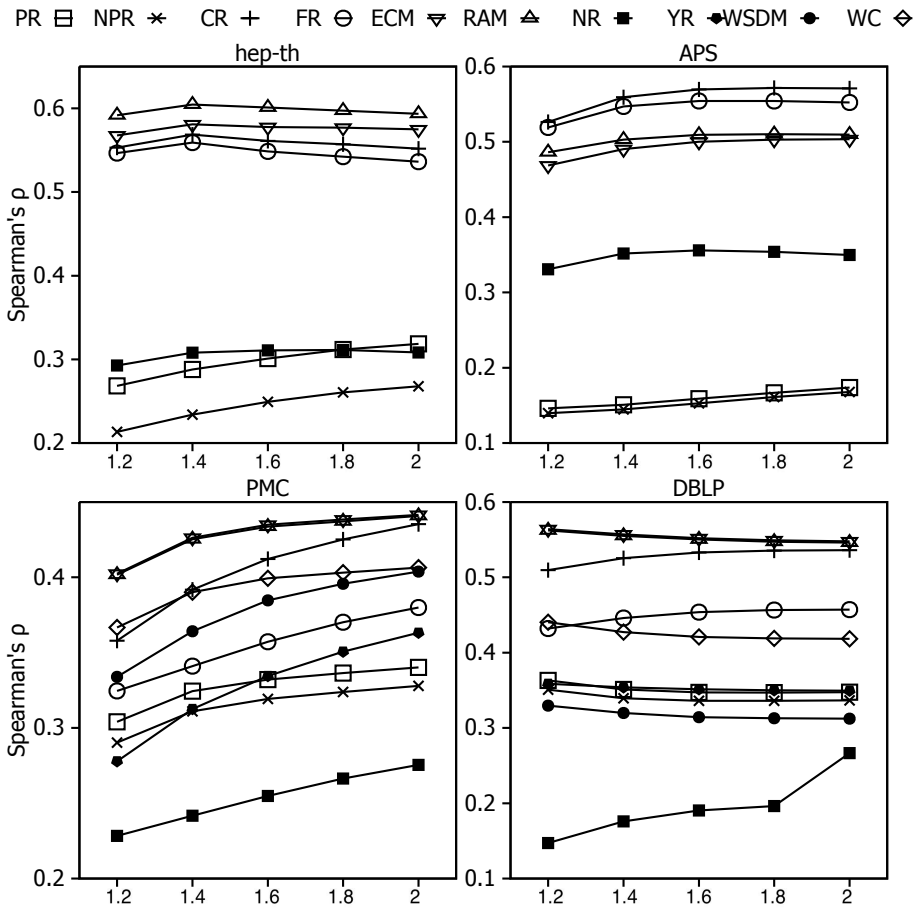
| **APS** | **P-CC** | | | | **P-PR** | | | |
|---|---|---|---|---|---|---|---|---|
| | $\rho$ | $\tau$ | **Ακρίβεια** | **nDCG** | $\rho$ | $\tau$ | **Ακρίβεια** | **nDCG** |
| PR | 0.159 | 0.113 | 0.160 | 0.347 | 0.127 | 0.085 | 0.120 | 0.347 |
| NPR | 0.153 | 0.109 | 0.220 | 0.359 | 0.134 | 0.090 | 0.180 | 0.378 |
| CR | **0.570** | **0.423** | 0.500 | 0.627 | **0.529** | **0.371** | 0.420 | 0.642 |
| FR | 0.554 | 0.412 | 0.540 | 0.658 | 0.518 | 0.361 | **0.440** | 0.653 |
| ECM | 0.500 | 0.377 | 0.540 | **0.716** | 0.399 | 0.280 | 0.420 | **0.672** |
| RAM | 0.509 | 0.385 | **0.580** | 0.705 | 0.412 | 0.289 | **0.440** | 0.667 |
| NR | 0.356 | 0.255 | 0.160 | 0.199 | 0.354 | 0.240 | 0.220 | 0.308 |

Πίνακας 4.8: PMC: μετρικές για τα P-CC, P-PR· $\eta = 1.6$, $k = 50$.

| **PMC** | **P-CC** | | | | **P-PR** | | | |
|---|---|---|---|---|---|---|---|---|
| | $\rho$ | $\tau$ | **Ακρίβεια** | **nDCG** | $\rho$ | $\tau$ | **Ακρίβεια** | **nDCG** |
| PR | 0.332 | 0.260 | 0.220 | 0.421 | 0.198 | 0.141 | 0.260 | 0.444 |
| NPR | 0.319 | 0.250 | 0.220 | 0.427 | 0.200 | 0.142 | 0.280 | 0.484 |
| CR | 0.412 | 0.316 | 0.360 | 0.648 | **0.272** | **0.189** | 0.440 | 0.717 |
| FR | 0.357 | 0.268 | 0.400 | 0.658 | 0.255 | 0.174 | 0.520 | **0.818** |
| ECM | **0.435** | **0.350** | 0.660 | 0.896 | 0.224 | 0.161 | 0.540 | 0.772 |
| RAM | 0.434 | **0.350** | **0.680** | **0.902** | 0.226 | 0.163 | **0.560** | 0.786 |
| NR | 0.255 | 0.193 | 0.300 | 0.482 | 0.245 | 0.170 | 0.520 | 0.767 |
| YR | 0.335 | 0.249 | 0.220 | 0.461 | 0.125 | 0.084 | 0.240 | 0.448 |
| WSDM | 0.385 | 0.291 | 0.140 | 0.382 | 0.041 | 0.027 | 0.140 | 0.317 |
| WC | 0.399 | 0.316 | 0.500 | 0.745 | 0.159 | 0.113 | 0.400 | 0.565 |

Πίνακας 4.9: DBLP: μετρικές για τα P-CC, P-PR· $\eta = 1.6$, $k = 50$.

| **DBLP** | **P-CC** | | | | **P-PR** | | | |
|---|---|---|---|---|---|---|---|---|
| | $\rho$ | $\tau$ | **Ακρίβεια** | **nDCG** | $\rho$ | $\tau$ | **Ακρίβεια** | **nDCG** |
| PR | 0.347 | 0.257 | 0.160 | 0.384 | 0.279 | 0.194 | 0.200 | 0.449 |
| NPR | 0.336 | 0.248 | 0.160 | 0.382 | 0.282 | 0.196 | 0.220 | 0.458 |
| CR | 0.533 | 0.397 | 0.440 | 0.717 | **0.496** | **0.348** | 0.500 | 0.765 |
| FR | 0.454 | 0.337 | 0.480 | 0.740 | 0.446 | 0.311 | 0.520 | 0.788 |
| ECM | **0.552** | **0.428** | **0.700** | **0.886** | 0.433 | 0.310 | **0.620** | **0.855** |
| RAM | 0.550 | 0.427 | 0.680 | 0.876 | 0.432 | 0.310 | **0.620** | 0.846 |
| NR | 0.262 | 0.190 | 0.300 | 0.561 | 0.310 | 0.212 | 0.420 | 0.684 |
| YR | 0.352 | 0.256 | 0.320 | 0.537 | 0.287 | 0.196 | 0.340 | 0.595 |
| WSDM | 0.314 | 0.227 | 0.320 | 0.442 | 0.145 | 0.097 | 0.300 | 0.464 |
| WC | 0.421 | 0.315 | 0.320 | 0.447 | 0.282 | 0.196 | 0.240 | 0.431 |

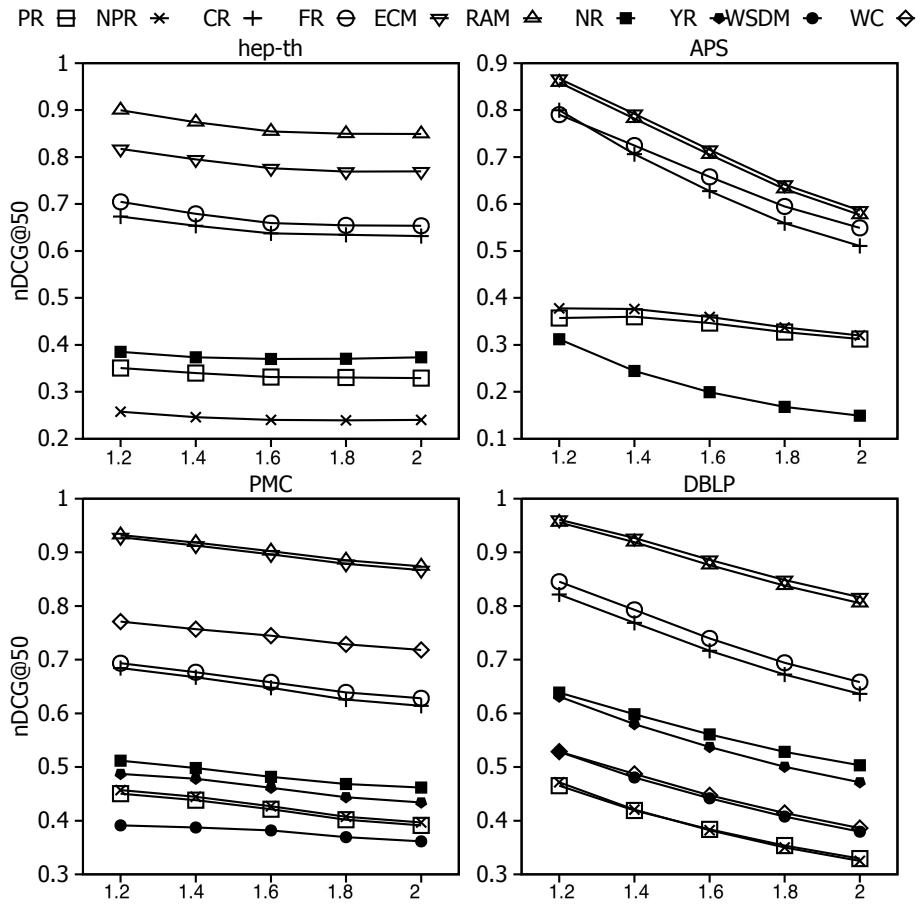



Σχήμα 4.4: Συσχέτιση της κατάταξης που παράγει κάθε μέθοδος με αυτή του υπόβαθρου αληθείας P-CC, μεταβάλλοντας το λόγο $\eta$.

ξεκάθαρα, για παράδειγμα, ότι οι μέθοδοι που δε χρησιμοποιούν χρονικούς παράγοντες (PR, NPR, WSDM) δεν είναι αποτελεσματικές σε αυτό το σενάριο. Ωστόσο, παρατηρούμε ότι δεν ευνοεί και η χρήση οποιουδήποτε τύπου χρονικών ποσοτήτων. Για παράδειγμα, η μέθοδος NR υπερβάλλει στη μεροληψία υπέρ των πιο πρόσφατων δημοσιεύσεων, καθώς χρησιμοποιεί τόσο την ηλικία της αναφερόμενης δημοσίευσης στον πίνακα μεταβάσεων όσο και πιθανότητες επιλογής που φθίνουν με την ηλικία των δημοσιεύσεων. Με τον τρόπο αυτό προσομοιώνει έναν ερευνητή που ξεκινάει την ανάγνωσή του από πρόσφατες δημοσιεύσεις, αλλά *και* προτιμάει να ακολουθεί αναφορές προς πρόσφατες δημοσιεύσεις. Από την άλλη, μέθοδοι όπως η WC προωθούν δημοσιεύσεις που αναφέρθηκαν σχετικά γρήγορα μετά την έκδοσή τους (χρήση του χρονικού διαστήματος αναφοράς), ανεξάρτητα από το αν αυτό συνέβη πρόσφατα, ή παλιότερα, και επομένως δεν ανταναμκλούν αναγκαστικά τις τρέχουσες δυναμικές του δικτύου αναφορών.

#### 4.3.2.2 Μεταβάλλοντας τον λόγο $\eta$

Σε αυτό το πείραμα μεταβάλλουμε το λόγο $\eta$ και μετράμε την αποτελεσματικότητα όλων των μεθόδων στην επίτευξη μιας κατάταξης που να συμφωνεί με το υπόβαθρο αληθείας P-CC. Παρουσιάζουμε τα αποτελέσματα των μετρήσεων του $\rho$ του Spearman για κάθε μέθοδο και κάθε σύνολο δεδομένων στο Σχήμα 4.4. Τα αποτελέσματα των μετρήσεων για το $\tau$ του Kendall είναι παρόμοια και γι' αυτό εδώ παραλείπονται. Οι

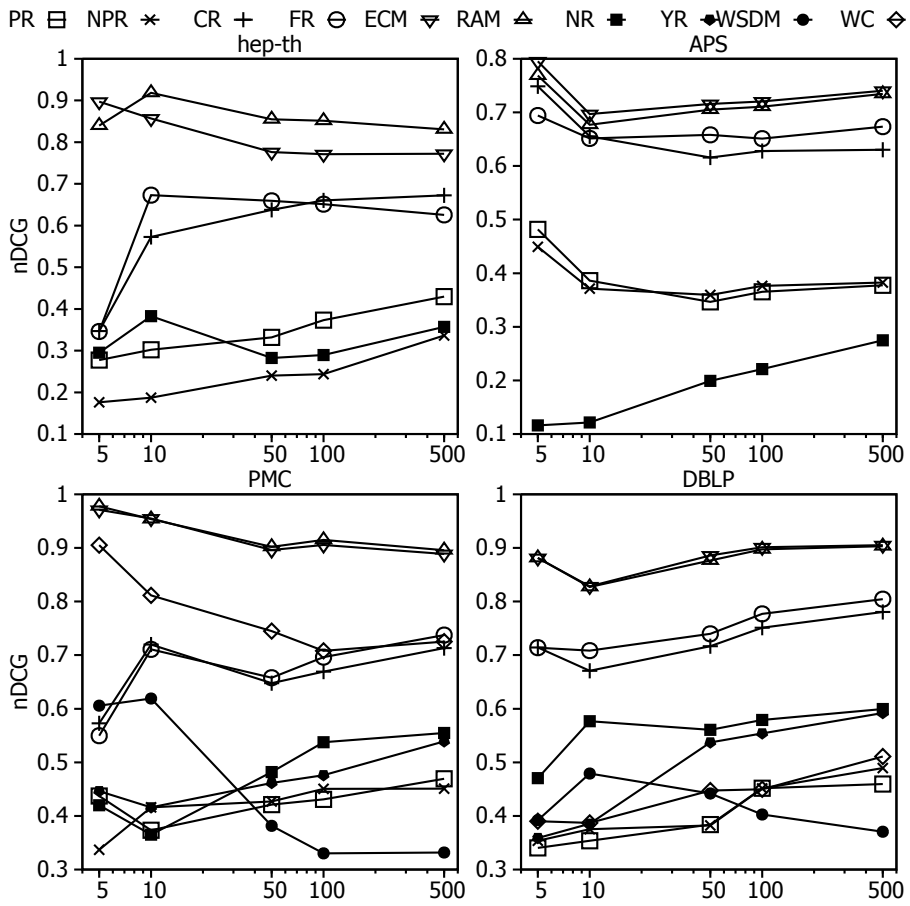



Σχήμα 4.5: nDCG@50 κάθε κατάταξης με αναφορά το υπόβαθρο αληθείας P-CC, μεταβάλλοντας το λόγο $\eta$.

γενικές παρατηρήσεις που έγιναν στην προηγούμενη υποενότητα ισχύουν και εδώ: οι μέθοδοι RAM, ECM, CR είναι οι πιο αποτελεσματικές σε όλα τα σύνολα δεδομένων, ανεξάρτητα από το $\eta$.

Καθώς αυξάνει η τιμή του $\eta$, παρατηρούμε ότι αρχικά η αποτελεσματικότητα των μεθόδων αυξάνει, έπειτα όμως φτάνει ένα άνω όριο και στη συνέχεια πέφτει. Αυτό οφείλεται σε δύο παράγοντες. Ο πρώτος παράγοντας σχετίζεται με το ακόλουθο γεγονός: οι μέθοδοι κάνουν κάποια ευθεία προέκταση των τάσεων των αναφορών του παρόντος και πρόσφατου παρελθόντος, με βάση την οποία ορίζουν την κατάταξη των δημοσιεύσεων στο μέλλον. Επομένως είναι φυσικό καθώς το $\eta$ αυξάνει, κάποια στιγμή η ακρίβεια αυτής της ευθείας προέκτασης να πέφτει. Γι' αυτό και για μεγάλα $\eta$ οι μέθοδοι πιάνουν κάποιο άνω όριο όσον αφορά τη συσχέτιση με το P-CC, ή και μειωμένες τιμές συσχέτισης.

Ο δεύτερος παράγοντας σχετίζεται με τη γενική κατανομή των αναφορών σε ένα δίκτυο αναφορών. Πρόκειται για κατανομή εκθετικού νόμου (Power Law), ή κάποια «συγγενική» κατανομή [11, 19]. Αυτό σημαίνει ότι η μεγάλη των πλειοψηφία δημοσιεύσεων λαμβάνει πολύ λίγες αναφορές συνολικά, σχηματίζοντας αυτό που είναι γνωστό ως «μεγάλη ουρά» (Long Tail) της κατανομής, ενώ οι δημοσιεύσεις με τις περισσότερες αναφορές σχηματίζουν την «κεφαλή» της κατανομής. Οι μικρές τιμές του $\eta$ αντιστοιχούν σε μικρή μελλοντική χρονική περίοδο. Συνεπώς, οι δημοσιεύσεις που βρίσκονται στην ουρά της κατανομής δεν προλαβαίνουν να συσσωρεύσουν τις αναφορές που απαιτούνται για να ξεχωρίσουν μεταξύ τους - οποιεσδήποτε διαφορές παρατηρούνται




Σχήμα 4.6: nDCG για την κατάταξη κάθε μεθόδου, με βάση το υπόβαθρο αληθείας P-CC, υπολογισμένο σε διαφορετικές του $k$· $\eta = 1.6$.

μεταξύ τους μπορεί να είναι τυχαίες. Ως εκ τούτου, η κατάταξή τους στο υπόβαθρο αληθείας βασίζεται σε ένα μικρό δείγμα αναφορών που δεν αντικατοπτρίζει πλήρως τη σχετική τους αξία. Επομένως όλες οι μέθοδοι αναμένεται να δυσκολεύονται να ξεχωρίσουν μεταξύ των δημοσιεύσεων που βρίσκονται στην ουρά της κατανομής. Επιπλέον, δεδομένου ότι οι δημοσιεύσεις στην ουρά είναι και οι πιο πολυπληθείς, αναμένουμε χαμηλές συσχετίσεις για μικρές τιμές του $\eta$. Με την αύξηση του $\eta$, το πρόβλημα αυτό διορθώνεται και η αποτελεσματικότητα όλων των μεθόδων αυξάνει.

Στο Σχήμα 4.5 παρουσιάζονται τα αποτελέσματα του nDCG ($k = 50$) όλων των μεθόδων καθώς μεταβάλλεται ο λόγος $\eta$. Η πιο ενδιαφέρουσα παρατήρηση που μπορεί να γίνει εδώ, είναι ότι όλες οι μέθοδοι δυσκολεύονται να εντοπίσουν τις πιο δημοφιλείς δημοσιεύσεις (με βάση τον αριθμό αναφορών), όσο κοιτάμε πιο μακριά στο μέλλον. Αυτό συμβαίνει λόγω της κατανομής που συζητήσαμε παραπάνω: δεδομένου ότι εδώ ενδιαφέρουν δημοσιεύσεις που λαμβάνουν μεγάλο αριθμό αναφορών, η ουρά της κατανομής δεν επηρεάζει. Συνεπώς, η κατανομή αναφορών των σημαντικότερων δημοσιεύσεων είναι εμφανής για μικρές τιμές του $\eta$. Με τη μεταβολή του $\eta$ φανερώνονται οι τάσεις διαφοροποίησης στη λήψη αναφορών μεταξύ των δημοσιεύσεων καθώς το $t_c + T$ απομακρύνεται από το $t_c$ και τότε η αποτελεσματικότητα των μεθόδων πέφτει.

#### 4.3.2.3 Μεταβάλλοντας το $k$

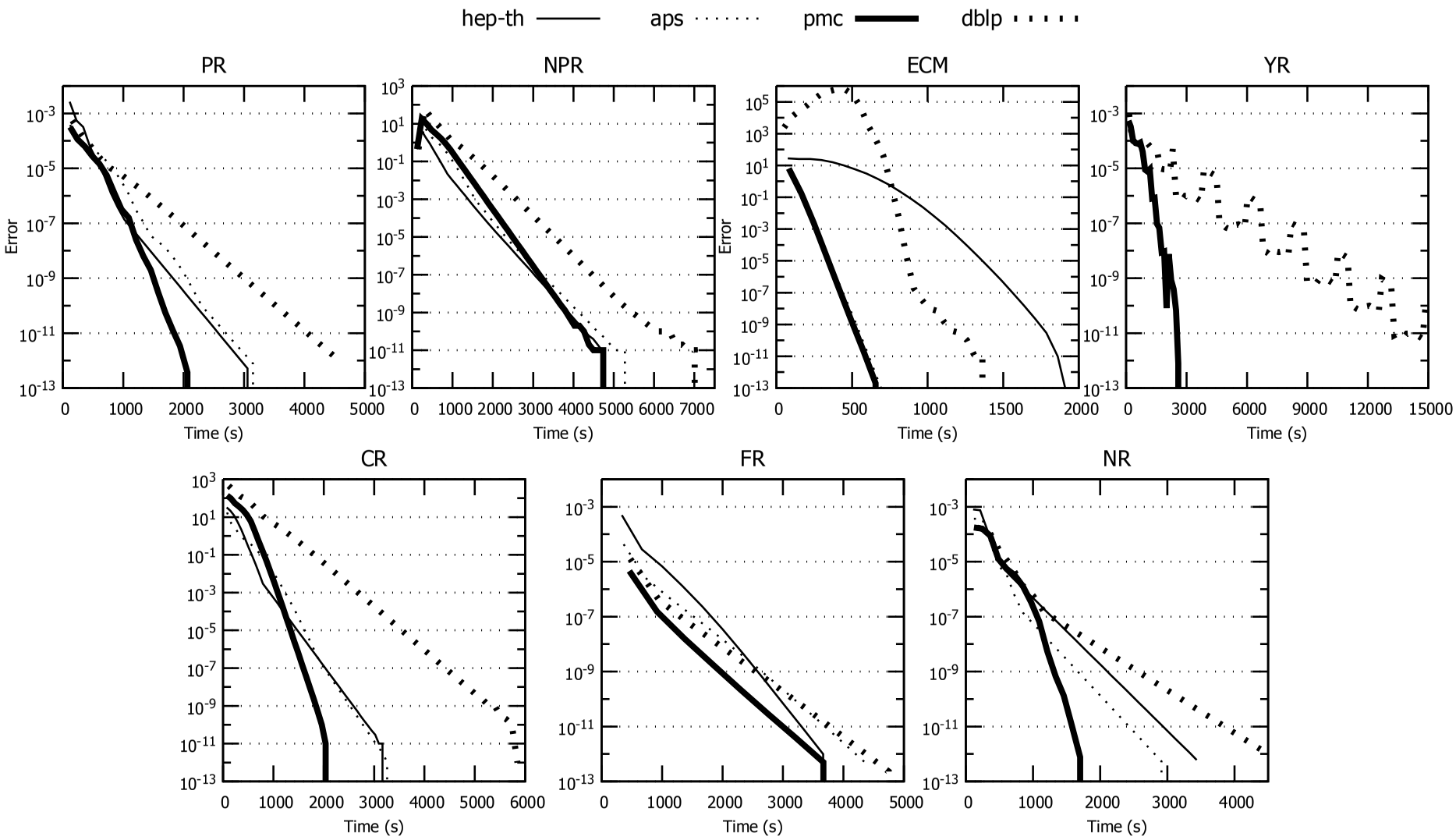
Στο τελευταίο πείραμα, μετράμε την τιμή nDCG κάθε μεθόδου, μεταβάλλοντας το $k$, με το $\eta$ να παραμένει σταθερό στην τιμή βάσης ($\eta = 1.6$). Στο Σχήμα 4.6 παρουσιάζουμε τα αποτελέσματα. Παρατηρούμε ότι οι μέθοδοι που ήταν αποτελεσματικές στο προηγούμενο πείραμα, οι RAM, ECM, παραμένουν αποτελεσματικές σε όλες τις τιμές του $k$ που εξετάζουμε. Αξίζει να σημειωθεί πως, παρόλο που οι μέθοδοι δεν μπορούν να παράξουν μια κατάταξη που να είναι συνολικά πολύ όμοια με το υπόβαθρο αληθείας P-CC, ωστόσο μπορούν να διακρίνουν ευκολότερα τις σημαντικότερες δημοσιεύσεις.

## 4.4 Σύγκλιση και Χρόνοι Εκτέλεσης

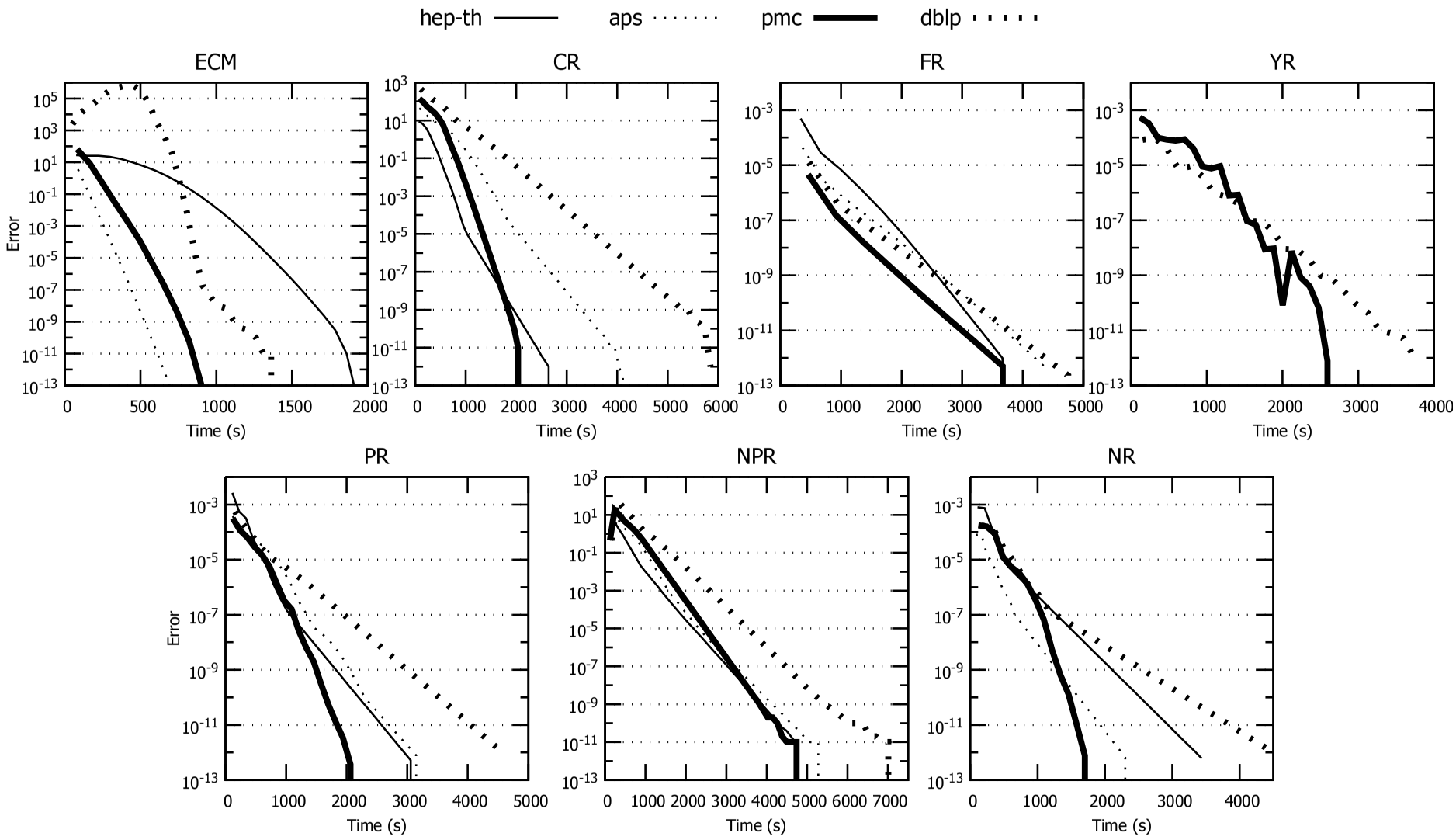
Σε αυτή την ενότητα συγκρίνουμε την ταχύτητα σύγκλισης των επαναληπτικών μεθόδων, θέτοντας τις παραμέτρους κάθε μεθόδου στις τιμές που πετυχαίνουν την καλύτερη συνολική συσχέτιση με τα υπόβαθρα αληθείας I-PR, P-CC, όπως παρουσιάστηκαν στις Ενότητες 4.3.1-4.3.2. Στα Σχήματα 4.7-4.8 παρουσιάζεται το σφάλμα σύγκλισης κάθε μεθόδου, για κάθε σύνολο δεδομένων, ως συνάρτηση του χρόνου, όταν χρησιμοποιούμε ως υπόβαθρα αληθείας τα I-PR, P-CC, αντίστοιχα. Σε κάθε σενάριο, παρουσιάζουμε τις μεθόδους σε σειρά αντίστοιχη με την αποτελεσματικότητά τους, βάσει της μετρικής $\rho$.

Σε όλες τις περιπτώσεις η μέθοδος ECM συγκλίνει ταχύτερα. Η μέθοδος αυτή βασίζεται στην κεντρικότητα Katz (βλ. Ενότητα 3.1.2) και επομένως βαθμολογεί κάθε κόμβο με βάση το σύνολο των μονοπατιών που περνάνε από αυτόν [51]. Η γρήγορη σύγκλιση της μεθόδου οφείλεται στις τιμές των παραμέτρων της, με βάση τις οποίες τα βάρη μονοπατιών μήκους μεγαλύτερο από 1 φθίνουν πολύ γρήγορα, με αποτέλεσμα η συνολική βαθμολογία κάθε κόμβου να υπολογίζεται σε λίγες επαναλήψεις. Η σύγκλιση του PR και των παραλλαγών του (NPR, NR, CR και YR) εξαρτάται από την πιθανότητα τυχαίας μετάβασης $\alpha$ (Εξίσωση 2.1), με τιμές της πιθανότητας κοντά στο 1 να συνεπάγονται έναν αυξανόμενο αριθμό επαναλήψεων για τη σύγκλιση (και, επομένως, και αυξημένο χρόνο εκτέλεσης) [47]. Παρατηρούμε ότι από αυτές τις μεθόδους, οι PR, CR, NR και YR συγκλίνουν σε συγκρίσιμο χρόνο. Αυτό συμβαίνει καθώς όλες χρησιμοποιούν μια τιμή $\alpha \sim 0.5$. Μια εξαίρεση σ' αυτό αποτελεί η περίπτωση του YR όταν χρησιμοποιούμε ως υπόβαθρο αληθείας το I-PR, όπου η μέθοδος YR συγκλίνει πολύ πιο αργά για το σύνολο δεδομένων του DBLP. Σε αυτό το σύνολο δεδομένων η μέθοδος πετυχαίνει την καλύτερη αποτελεσματικότητά της για $\alpha = 0.85$, αντί για $\alpha = 0.5$ που την κάνει πιο αποτελεσματική στο σύνολο PMC. Παρουσιάζει επιπλέον ενδιαφέρον το γεγονός ότι η μέθοδος NPR που αποτελεί απλή παραλλαγή του απλού PR, με παρόμοια τιμή του $\alpha$ συγκλίνει αισθητά πιο αργά, εξαιτίας των μη γραμμικών υπολογισμών που εκτελεί σε κάθε επανάληψη. Συγκεκριμένα, η μέθοδος NPR απαιτεί περίπου $15 - 25$ περισσότερες επαναλήψεις σε σύγκριση με το απλό PR, από όπου προκύπτει η πιο αργή σύγκλιση. Τέλος, η μέθοδος FR συγκλίνει αισθητά πιο αργά σε σχέση με το απλό PR. Αξίζει να σημειωθεί ότι αυτό δε σχετίζεται με τον απαιτούμενο αριθμό επαναλήψεων της μεθόδου, αλλά με το γεγονός ότι η μέθοδος απαιτεί σημαντικά περισσότερους υπολογισμούς ανά επανάληψη, καθώς χρησιμοποιεί πολλαπλά δίκτυα.

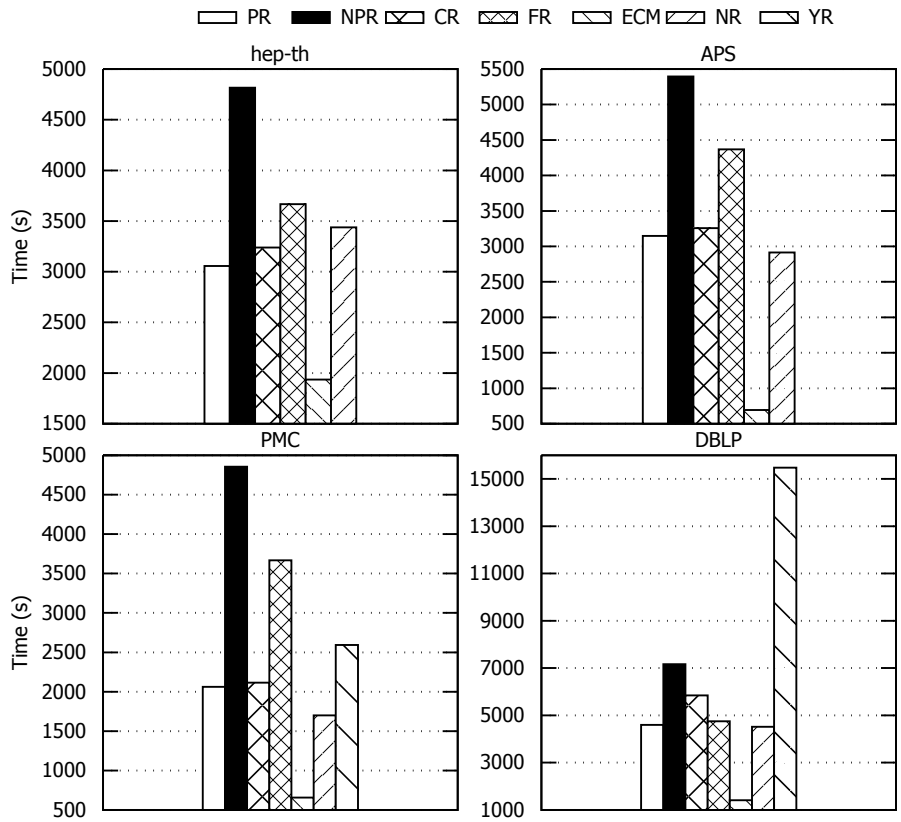
Για να γίνει πιο σαφής η σύγκριση, παραθέτουμε και τους συνολικούς χρόνους εκτέλεσης (έως τη σύγκλιση) κάθε μεθόδου, με σφάλμα σύγκλισης $\epsilon < 10^{-12}$, στα Σχήματα 4.9 και 4.10, για τα σενάρια όπου χρησιμοποιούμε ως υπόβαθρα αληθείας τα




Σχήμα 4.7: Ταχύτητες σύγκλισης όλων των μεθόδων σε κάθε σύνολο δεδομένων, βάσει του συνόλου παραμέτρων που οδηγεί στην καλύτερη αποτελεσματικότητα όταν χρησιμοποιούμε ως υπόβαθρο αληθείας το I-PR.




Σχήμα 4.8: Ταχύτητες σύγκλισης όλων των μεθόδων σε κάθε σύνολο δεδομένων, βάσει του συνόλου παραμέτρων που οδηγεί στην καλύτερη αποτελεσματικότητα όταν χρησιμοποιούμε ως υπόβαθρο αληθείας το P-CC.



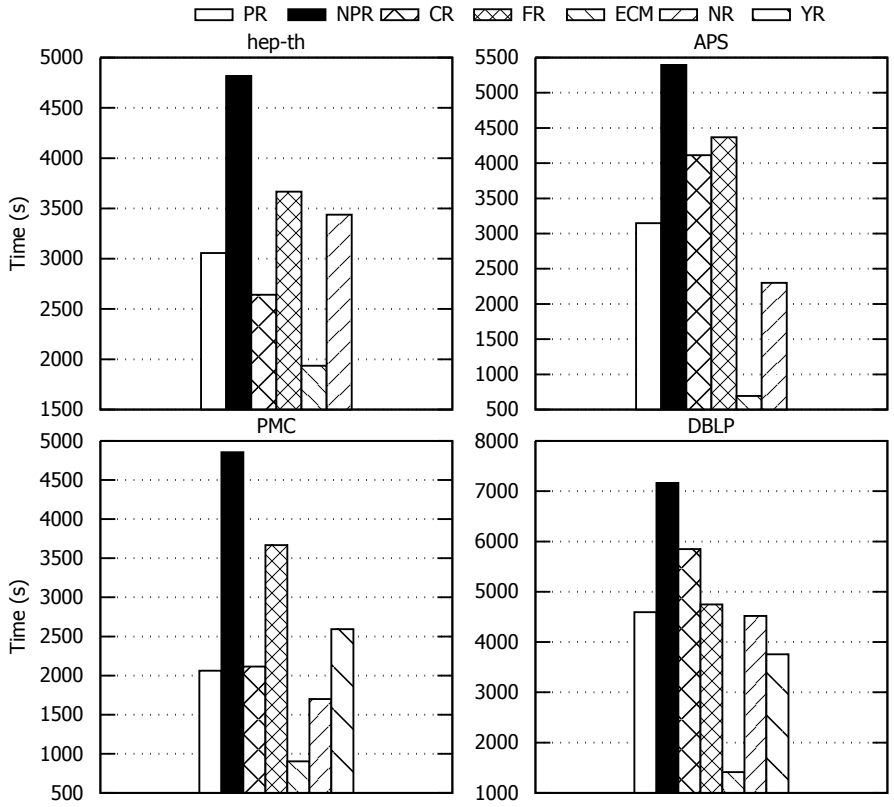

Σχήμα 4.9: Χρόνοι εκτέλεσης ανά μέθοδο, βάσει των παραμέτρων που οδηγούν στην καλύτερη αποτελεσματικότητα στο σενάριο I-PR.

I-PR και P-CC, αντίστοιχα. Όταν υπάρχουν περιορισμοί χρόνου, τότε μπορούμε να θεωρήσουμε προτιμητέα τη χρήση του PR στο σενάριο της επιρροής, καθώς είναι αρκετά γρήγορο (μόνο η μέθοδος ECM είναι ταχύτερη) και έχει τη μεγαλύτερη αποτελεσματικότητα. Στην περίπτωση του P-CC η μέθοδος ECM είναι ξεκάθαρα η καλύτερη επιλογή, καθώς είναι ταυτόχρονα η πιο γρήγορη και πιο αποτελεσματική.

## 4.5 Συμπεράσματα

Η αξιολόγηση που πραγματοποιήσαμε βασίζεται σε τέσσερα πραγματικά σύνολα δεδομένων και εστίασε σε τρία ερευνητικά ερωτήματα: (1) πόσο διακριτές είναι οι έννοιες της μακροχρόνιας και βραχυχρόνιας απήχησης (επιρροής και δημοφιλίας, αντίστοιχα), (2) ποια μέθοδος αποτελεί την τεχνολογία αιχμής για την κατάταξη με βάση το κάθε είδος απήχησης, (3) ποια (επαναληπτική) μέθοδος συγκλίνει ταχύτερα. Σε όσα ακολουθούν, συνοψίζουμε τα συμπεράσματά μας σχετικά με αυτά τα ερωτήματα και καταλήγουμε σε κάποιες γενικές παρατηρήσεις για την τρέχουσα κατάσταση της έρευνας πάνω στο αντικείμενο.

Όσον αφορά το πρώτο ερώτημα, παρατηρούμε ότι η δημοφιλία και η επιρροή αντανακλούν διακριτές ιδιότητες, που όμως συσχετίζονται σε κάποιο βαθμό. Οι συσχετίσεις μεταξύ των ειδών δημοφιλίας και επιρροής, που ορίζονται με χρήση του βαθμού αναφορών και του PageRank, είναι ισχυρότερες όταν αφορούν στο ίδιο είδος απήχησης (στα ζευγάρια P-CC/P-PR και I-CC/I-PR). Επιπλέον, όσο περισσότερο κοιτάμε στο μέλλον (μεγάλες τιμές του $\eta$), τόσο η ομοιότητα των κατατάξεων βάσει δημοφιλίας και επιρροής αυξάνει. Αυτή η παρατήρηση οδηγεί στο ζήτημα της ορθής επιλογής




Σχήμα 4.10: Χρόνοι εκτέλεσης ανά μέθοδο, βάσει των παραμέτρων που οδηγούν στην καλύτερη αποτελεσματικότητα στο σενάριο P-CC.

του χρονικού ορίζοντα $T$ (ή αλλιώς του λόγου $\eta$), έτσι ώστε το υπόβαθρο αληθείας να ανταναμλά πράγματι την κατάταξη με βάση την τρέχουσα δημοφιλία των δημοσιεύσεων. Ο χρονικός ορίζοντας αυτός, θα πρέπει αφενός να είναι μικρός, ώστε να μη γίνεται σύγχυση βραχυχρόνιας και μακροχρόνιας απήχησης, αφετέρου να είναι αρκετά μεγάλος, ώστε να περιλαμβάνει την τυπική διάρκεια του κύκλου έρευνας του εκάστοτε επιστημονικού πεδίου. Σημειώνουμε επίσης, ότι τα υπόβαθρα αληθείας που αφορούν τη μακροχρόνια απήχηση (I-CC και I-PR) είναι μεταξύ τους ισχυρά συσχετισμένα για μεγάλες τιμές του λόγου $\eta$. Από αυτό υπονοείται πως είναι λογικό κατά την βελτιστοποίηση των παραμέτρων μιας μεθόδου να εστιάζει κανείς σε ένα από τα δύο μέτρα κάθε φορά, π.χ., στο I-PR, όπως συμβαίνει στη τρέχουσα βιβλιογραφία.

Σε σχέση με το δεύτερο ερευνητικό ερώτημα, δεν μπορούμε να εντοπίσουμε μια μοναδική μέθοδο που να είναι η αποτελεσματικότερη τόσο στην παραγωγή κατατάξεων βάσει της δημοφιλίας, όσο και της επιρροής. Αντιθέτως, παρατηρούμε ότι σε πολλές περιπτώσεις η επιλογή της αποτελεσματικότερης μεθόδου εξαρτάται από το πώς ορίζεται η απήχηση, καθώς και στον εκάστοτε στόχο, λ.χ., αν στοχεύουμε στην παραγωγή μιας συνολικής κατάταξης, ή στη σωστή κατάταξη των πιο σημαντικών δημοσιεύσεων με βάση την απήχηση. Στον Πίνακα 4.10 συνοψίζουμε τα συμπεράσματα μας σχετικά με το σενάριο στο οποίο είναι αποτελεσματικότερη η κάθε μέθοδος.

Στην περίπτωση της επιρροής μπορούμε να συνοψίσουμε τα ακόλουθα συμπεράσματα. Γενικά η χρήση χρονικά ενήμερων πινάκων γειτνίασης είναι ένας καλός μηχανισμός για την κατάταξη των σημαντικότερων δημοσιεύσεων με βάση την επιρροή. Από την άλλη, η χρήση πιθανοτήτων επιλογής που φθίνουν με την ηλικία των δημοσιεύσεων εμπεριέχει κάποιες αντιφάσεις. Από τη μια μειώνει τη συνολική συσχέτιση με το υ-

Πίνακας 4.10: Τεχνολογίες Αιχμής στις περιπτώσεις της παραγωγής συνολικής κατάταξης/προσδιορισμού των $k$ σημαντικότερων αποτελεσμάτων.

| | **Επιρροή** | **Δημοφιλία** |
|---|---|---|
| **Αριθμός Αναφορών** | RAM/RAM | ECM/RAM |
| **PageRank** | PR/CR | CR/RAM |

πόβαθρο αληθείας, αλλά από την άλλη ευνοεί τον εντοπισμό των σημαντικότερων δημοσιεύσεων βάσει την επιρροής, ειδικά όταν αυτή ορίζεται μέσω του PR. Αυτός ο τρόπος ορισμού της επιρροής είναι και ο συχνότερα χρησιμοποιούμενος στη βιβλιογραφία. Το απλό PageRank είναι αποτελεσματικό όσον αφορά τη παραγωγή συνολικής κατάταξης. Η χρήση μεταδεδομένων, πολλαπλών δικτύων και συνδυαστικών μεθόδων δε φαίνεται να οδηγεί σε καλύτερα αποτελέσματα. Πρέπει να σημειωθεί επιπλέον ότι, όπως αναμένεται, οι μέθοδοι που βασίζονται σε παραλλαγές του κάθε είδους κεντρικότητας που έχουμε εξετάσει (αριθμός αναφορών, PageRank) είναι αποτελεσματικότερες όταν η επιρροή ορίζεται με χρήση του αντίστοιχου μέτρου κεντρικότητας.

Όσον αφορά τη δημοφιλία, τα συμπεράσματά μας είναι τα εξής. Η χρήση χρονικών παραγόντων είναι ζωτικής σημασίας για τη παραγωγή κατατάξεων βάσει της δημοφιλίας, καθώς αντισταθμίζουν την υπέρ των παλαιότερων δημοσιεύσεων. Συνολικά, στο σενάριο της δημοφιλίας, φαίνεται ότι όπως και στο σενάριο της επιρροής, η χρήση μεταδεδομένων, πολλαπλών δικτύων και συνδυαστικών μεθόδων δεν οδηγεί σε καλύτερα αποτελέσματα. Σημειώνουμε επίσης ότι η κατάταξη βάσει της δημοφιλίας φαίνεται να αποτελεί ένα δυσκολότερο πρόβλημα απ' ότι η κατάταξη βάσει της επιρροής. Η αποτελεσματικότητα όλων των μεθόδων είναι χαμηλότερη, ειδικά στα πειράματα με μεγαλύτερους χρονικούς ορίζοντες (μεγαλύτερες τιμές του $\eta$). Για παράδειγμα, παρατηρούμε ότι η μεγαλύτερη ακρίβεια στα 50 σημαντικότερα αποτελέσματα στην περίπτωση της δημοφιλίας είναι 0.7 ενώ στην περίπτωση της επιρροής είναι 0.9. Η παρατήρηση αυτή εξηγεί γιατί η βιβλιογραφία εστιάζει περισσότερο στην αποτελεσματικότητα των μεθόδων ως προς τη παραγωγή κατατάξεων βάσει της δημοφιλίας. Οι τρέχουσες τεχνολογίες αιχμής αφήνουν ανοιχτά περιθώρια βελτίωσης σ' αυτή τη κατεύθυνση.

Τέλος, όσον αφορά το τρίτο ερευνητικό ερώτημα που θέσαμε, συμπεραίνουμε ότι μεταξύ των μεθόδων PR και CR που είναι οι πιο αποτελεσματικές στο σενάριο της επιρροής, η μέθοδος PR μπορεί να είναι προτιμητέα όταν υπάρχουν περιορισμοί στο διαθέσιμο χρόνο υπολογισμού, καθώς συγκλίνει γρηγορότερα σε όλες τις περιπτώσεις. Στην περίπτωση της δημοφιλίας, όπου οι μέθοδοι CR και RAM/ECM είναι οι πιο αποτελεσματικές, οι RAM/ECM μπορεί να είναι προτιμότερες σε περιπτώσεις όπου υπάρχουν χρονικοί περιορισμοί.

# Κεφάλαιο 5

# Αποτελεσματική Κατάταξη Δημοσιεύσεων Βάσει Απήχησης

Σε αυτό το κεφάλαιο εξετάζουμε το ζήτημα της κατάταξης δημοσιεύσεων επεκτείνοντας θεωρητικά και πειραματικά τα συμπεράσματα για τις μεθόδους κατάταξης που εξετάσαμε στο Κεφάλαιο 4. Συγκεκριμένα, στην υποενότητα 5.1 περιγράφουμε δύο πλήρως λειτουργικά και αποτελεσματικά συστήματα αναζήτησης επιστημονικής βιβλιογραφίας που αναπτύξαμε, τα οποία κάνουν χρήση μεθόδων κατάταξης: το BIP! Finder και το mirPub v2. Στην υποενότητα 5.2, με αφετηρία τα περιθώρια βελτίωσης στο σενάριο κατάταξης με βάση τη δημοφιλία, τα οποία παρατηρήσαμε στο Κεφάλαιο 4, σχεδιάζουμε και υλοποιούμε μια νέα μέθοδο που παράγει βελτιωμένες κατατάξεις ως προς αυτό το είδος απήχησης, σε σχέση με τη τρέχουσα βιβλιογραφία. Στη συνέχεια αξιολογούμε πειραματικά τη μέθοδο μας, δείχνοντας ότι υπερισχύει έναντι των άλλων μεθόδων στο σενάριο κατάταξης με βάση τη δημοφιλία.

## 5.1 Τεχνολογίες Κατάταξης Βάσει Απήχησης σε Μηχανές Αναζήτησης

Παρουσιάζουμε σε αυτή την ενότητα δύο μηχανές αναζήτησης επιστημονικής βιβλιογραφίας: το BIP! Finder (στο εξής BIP!) και το mirPub v2 (στο εξής mirPub), η ανάπτυξη των οποίων έγινε στα πλαίσια της διατριβής. Πρόκειται για μηχανές αναζήτησης βιβλιογραφίας γενικού και εξειδικευμένου σκοπού, αντίστοιχα. Το BIP! είναι μια μηχανή αναζήτησης επιστημονικής βιβλιογραφίας γενικού σκοπού, η οποία βασίζεται στα δεδομένα αναφορών μεταξύ δημοσιεύσεων που παρέχονται από την πρωτοβουλία Open Citations[1] και στη συλλογή μεταδεδομένων για δημοσιεύσεις από το Open Academic Graph.[2] Η μηχανή mirPub, από την άλλη, αφορά σε αναζήτηση βιβλιογραφίας από τις επιστήμες ζωής, συγκεκριμένα δημοσιεύσεις σχετικές με βιομόρια microRNA (miRNA).

[1] https://opencitations.net/
[2] https://www.openacademic.ai/

### 5.1.1 BIP! Finder

Η μηχανή αναζήτησης BIP! Finder[3] (BIbliograPhy Finder) αναπτύχθηκε με αφορμή το γεγονός ότι ο μεγάλος όγκος επιστημονικών δημοσιεύσεων δημιουργεί δυσκολίες στους ερευνητές, οι οποίοι θέλουν να ξεχωρίσουν εύκολα και γρήγορα τις πιο σημαντικές ανάμεσά τους. Για να αντιμετωπίσει το πρόβλημα αυτό, το BIP! στηρίζεται στη χρήση μεθόδων κατάταξης δημοσιεύσεων με βάση την απήχησή τους. Επιπλέον συνδυάζει τις βαθμολογίες απήχησης μαζί με παραδοσιακές βαθμολογίες κατάταξης που σχετίζονται με το εκάστοτε ερώτημα, δηλαδή βαθμολογίες που προκύπτουν από κλασσικές μεθόδους που χρησιμοποιούνται στην ανάκτηση πληροφορίας (Information Retrieval). Οι περισσότερες μηχανές αναζήτησης επιστημονικής βιβλιογραφίας (π.χ., Google Scholar, CiteSeer$^x$) χρησιμοποιούν είτε μια τέτοια παραδοσιακή προσέγγιση, είτε στηρίζονται στον αριθμό αναφορών για να καθορίσουν την σειρά εμφάνισης των αποτελεσμάτων αναζήτησης στο χρήστη.

Σημαντική καινοτομία της συγκεκριμένης μηχανής αναζήτησης, είναι ότι αποφεύγει την εσφαλμένη θεώρηση ότι υπάρχει ένα μέτρο επαρκές για να κατατάξει τις δημοσιεύσεις καλύπτοντας όλες τις ανάγκες αναζήτησης. Αντιθέτως, θεωρεί αυτή τη προσέγγιση μια υπεραπλούστευση, πράγμα που όπως δείξαμε και στο Κεφάλαιο 4 ισχύει, δεδομένου ότι μπορούμε να διακρίνουμε τουλάχιστον δύο είδη απήχησης: τη δημοφιλία και την επιρροή. Αυτό έχει ιδιαίτερη σημασία στα πλαίσια μιας μηχανής αναζήτησης βιβλιογραφίας, καθώς το κάθε είδος απήχησης σχετίζεται και με διαφορετικές ανάγκες χρηστών.

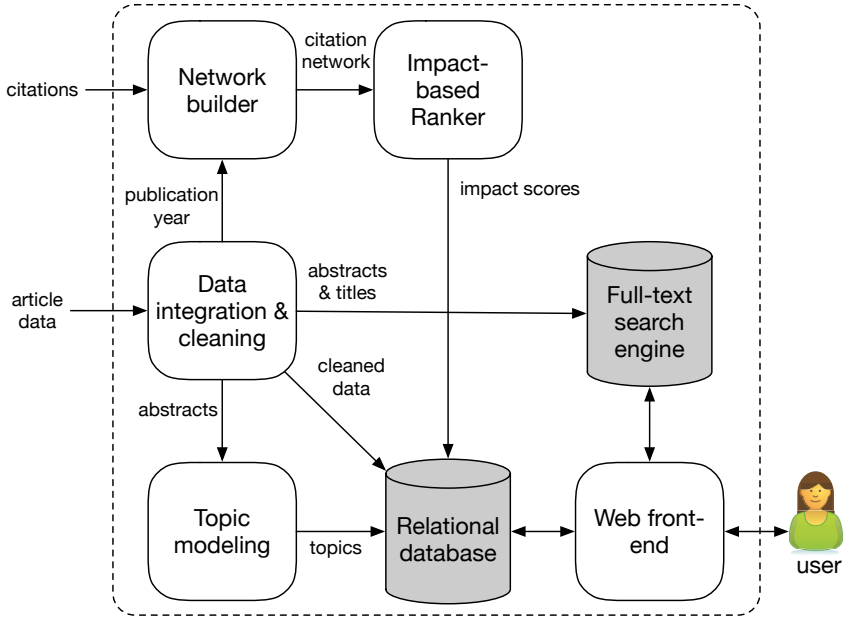
Για παράδειγμα, ένας έμπειρος ερευνητής, με καλή γνώση του ερευνητικού του αντικειμένου, είναι βέβαιο ότι γνωρίζει τις σημαντικότερες εργασίες του χώρου που μελετάει. Ωστόσο, χρειάζεται να έχει εποπτεία των εξελίξεων στο χώρο και έτσι, όταν αναζητάει δημοσιεύσεις αυτό που τον ενδιαφέρει περισσότερο είναι η δημοφιλία τους. Από την άλλη, ένας νέος ερευνητής που ασχολείται για πρώτη φορά με το αντικείμενό του είναι πιθανότερο ότι θα πρέπει να ξεκινήσει μελετώντας τις δημοσιεύσεις εκείνες που διαμόρφωσαν τον ερευνητικό του χώρο και άρα τον ενδιαφέρει η επιρροή τους. Οι τρέχουσες μηχανές αναζήτησης (που βασίζονται στον αριθμό αναφορών ή/και στη σχετικότητα των κειμένων με το ερώτημα αναζήτησης), θα ικανοποιούσαν εν μέρει τον δεύτερο ερευνητή, ωστόσο θα υστερούσαν σημαντικά στην ικανοποίηση των αναγκών του πρώτου. Αυτό συμβαίνει γιατί η χρήση του αριθμού αναφορών είναι μεροληπτική υπέρ των παλιότερων δημοσιεύσεων [15], καθώς οι νεότερες χρειάζονται μήνες ή και χρόνια για να λάβουν έστω και τις πρώτες τους αναφορές [31].

Βάσει της συζήτησης που αναπτύχθηκε στην παρούσα ενότητα, το σύστημα BIP! αντιμετωπίζει μια σειρά ζητημάτων που αφορούν τη διαδικασία αναζήτησης βιβλιογραφίας. Οι συνεισφορές του συστήματος συνοψίζονται ως εξής:

- Υποστηρίζει κατάταξη αποτελεσμάτων συνδυάζοντας τη σχετικότητα της αναζήτησης με τη δημοφιλία/επιρροή, ενώ παράλληλα προσφέρει πολλά φίλτρα που εφαρμόζουν στα αποτελέσματα αναζήτησης. Στηρίζεται στις μεθόδους PageRank [59] και RAM [28], για την κατάταξη αποτελεσμάτων βάση της επιρροής και της δημοφιλίας τους, αντίστοιχα.

- Το BIP! παρέχει ένα ελεύθερα προσβάσιμο API από το οποίο μπορούν να αντληθούν οι βαθμολογίες κατάταξης που έχουν υπολογιστεί για τις δημοσιεύσεις του συστήματος. Με αυτόν τον τρόπο δίνει τη δυνατότητα, αλλά και κίνητρο για την

[3] https://bip.imsi.athenarc.gr



Σχήμα 5.1: Η αρχιτεκτονική του BIP! Finder.

ανάπτυξη εφαρμογών από τρίτους, δίνοντας προστιθέμενη αξία στην αγορά της ερευνητικής αναλυτικής.

- Παρέχει μια σειρά οπτικοποιήσεων που χρησιμεύουν στην καλύτερη κατανόηση των χαρακτηριστικών κάθε επιστημονικής δημοσίευσης (π.χ., την απήχηση, τα θέματα της δημοσίευσης κλπ.), καθώς και επιπλέον λειτουγρικότητες όπως η οργάνωση δημοσιεύσεων σε φακέλους με «αγαπημένα» (Bookmarks).

#### 5.1.1.1 Αρχιτεκτονική

Οι λειτουργικότητα του BIP! στηρίζεται σε μια σειρά από λογισμικά που αναπτύχθηκαν. Το Σχήμα 5.1 συνοψίζει την αρχιτεκτονική του, περιγράφοντας αυτά τα λογισμικά και τη ροή δεδομένων μεταξύ τους. Στις ακόλουθες παραγράφους τα περιγράφουμε αναλυτικότερα.

**Σύστημα Κατασκευής Δικτύου (Network Builder).** Αυτό το λογισμικό είναι υπεύθυνο για το «χτίσιμο» του δικτύου αναφορών που χρησιμοποιεί το σύστημα. Σαν είσοδό του παίρνει την τελευταία έκδοση του συνόλου δεδομένων COCI[4] της OpenCitations, που περιέχει περίπου 450 εκατομμύρια αναφορές μεταξύ 45 εκατομμυρίων άρθρων. Σημειώνεται εδώ ότι το σύνολο δεδομένων COCI περιέχει πληροφορία που σχετίζεται με ψηφιακά αναγνωριστικά DOI. Συνεπώς, κάποιες επιπλέον πληροφορίες για τις δημοσιεύσεις, που απαιτούνται από κάποιες μεθόδους κατάταξης (π.χ., το έτος δημοσίευσης) αναγκαστικά συλλέγονται από άλλα λογισμικά του συστήματος.

**Σύστημα Ολοκλήρωσης και Καθαρισμού Δεδομένων (Data integration & cleaning component.)** Αυτό το λογισμικό συλλέγει τα δεδομένα των δημοσιεύσεων (π.χ., τίτλους, περιλήψεις, συγγραφείς, περιοδικά, ημερομηνίες δημοσίευσης) από πολλαπλές πηγές. Στη παρούσα φάση, το BIP! συλλέγει δεδομένα από το REST API της Crossref[5] και το Open Academic Graph[6] [69, 73]. Εξαιτίας της χρήσης πολ-

[4]http://opencitations.net/download
[5]https://www.crossref.org/services/metadata-delivery/rest-api/
[6]https://www.openacademic.ai/oag/

λαπλών πηγών, ενδέχεται να υπάρχουν ασυνέπειες, ή επαναλήψεις και επικαλύψεις στα δεδομένα. Για το λόγο αυτό λαμβάνει χώρα ο καθαρισμός και η ολοκλήρωση των δεδομένων. Για παράδειγμα, το καθάρισμα των ονομάτων των περιοδικών και συνεδρίων περιλαμβάνει μεταξύ των άλλων τα ακόλουθα: αφαίρεση περιττών χαρακτήρων κενού, εφαρμογή κανόνων για μετατροπή του αρχικού γράμματος κάποιων λέξεων σε κεφαλαίο, χειρισμός παραλλαγών ονομάτων (π.χ., αφαίρεση της αρίθμησης στο όνομα ενός συνεδρίου), κλπ. Παρόμοιες διαδικασίες ακολουθούνται και στον καθαρισμό άλλων δεδομένων των δημοσιεύσεων, όπως τα ονόματα των συγγραφέων. Η έξοδος που παράγεται από το λογισμικό αποθηκεύεται σε σχεσιακή βάση δεδομένων του συστήματος και χρησιμοποιείται από τη διαδικτυακή διεπαφή για να παράξει το μεγαλύτερο μέρος του δυναμικού περιεχομένου που προβάλλεται στο χρήστη.

**Σύστημα Κατάταξης με Βάση την Απήχηση (Impact-based Ranker).** Αυτό το λογισμικό υλοποιεί τις μεθόδους κατάταξης που χρησιμοποιεί το BIP!. Οι αλγόριθμοι που χρησιμοποιούνται, όπως αναφέρεται παραπάνω, είναι οι PageRank και RAM (Κεφάλαιο 3). Συγκεκριμένα χρησιμοποιούνται οι Map-Reduce υλοποιήσεις των ελεύθερα διαθέσιμων μεθόδων που υλοποιήσαμε και τις οποίες περιγράψαμε στο Κεφάλαιο 4.

**Λογισμικό Μοντελοποίησης Θεμάτων (Topic Modelling Component).** Αυτό το λογισμικό παίρνει ως είσοδο τις περιλήψεις των άρθρων και τις χρησιμοποιεί για να εκπαιδεύσει ένα μοντέλο LDA [7] για την εξαγωγή λανθάνοντων θεμάτων κάθε δημοσίευσης. Για την εκπαίδευση του μοντέλου χρησιμοποιήθηκε η βιβλιοθήκη μοντελοποίησης θεμάτων gensim.[7] Με αυτή τη βιβλιοθήκη εκπαιδεύτηκε ένα μοντέλο για 500 θέματα. Στη συνέχεια, εξήχθησαν για κάθε άρθρο τα 3 σημαντικότερα θέματα που το περιγράφουν και αποθηκεύτηκαν στη σχεσιακή βάση του BIP!.

**Δικτυακή Διεπαφή, Αποθήκευση Δεδομένων & Ευρετηριοποίηση.** Η διαδικτυακή διεπαφή του BIP! υλοποιήθηκε σε γλώσσα PHP με χρήση της αρχιτεκτονικής MVC. Όλες οι οπτικοποιήσεις υλοποιήθηκαν με χρήση CSS και JavaScript, χρησιμοποιώντας επιπλέον βιβλιοθήκες (π.χ., D3.js). Όλα τα δεδομένα αποθηκεύονται σε σχεσιακή βάση δεδομένων. Επιπλέον, οι τίτλοι και οι περιλήψεις των δημοσιεύσεων αποθηκεύονται και ευρετηριοποιούνται με τη μηχανή αναζήτησης για κείμενα Solr[8] της Apache, η οποία τρέχει σε μια συστάδα από 3 εικονικά μηχανήματα (Virtual Machines - VMs), με 8 πυρήνες και 16GB RAM ανά κόμβο.

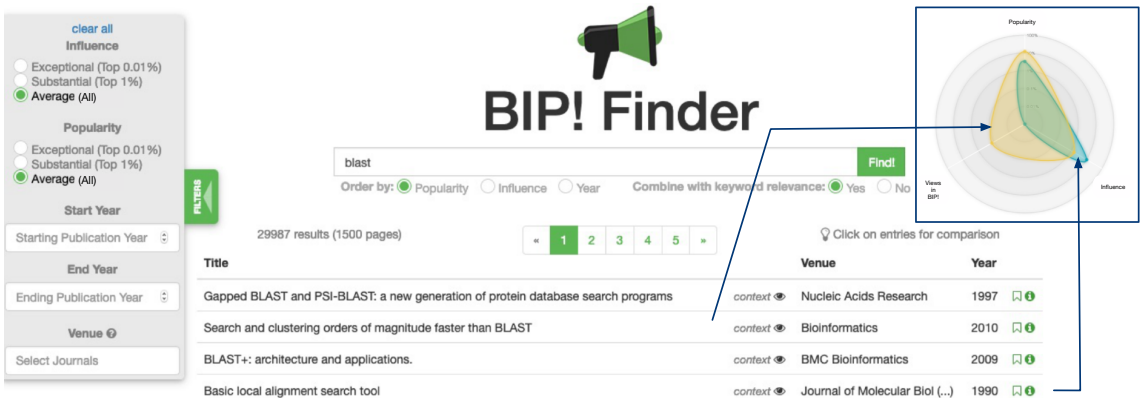
#### 5.1.1.2 Διεπαφή Χρήστη και Λειτουργίες

Το σύστημα BIP! παρέχει μια σειρά από λειτουργίες τόσο μέσω της διεπαφής χρήστη, όσο και μέσω του ελεύθερου API που διαθέτει. Περιγράφουμε εδώ τις επιμέρους λειτουργικότητες του συστήματος.

**Αναζήτηση Δημοσιεύσεων.** Πυρήνας του συστήματος είναι η μηχανή αναζήτησής του, που βασίζεται στην εισαγωγή λέξεων-κλειδιών από το χρήστη. Ο χρήστης εισάγει τις λέξεις-κλειδιά και αφού πατήσει το κουμπί "Find" του επιστρέφονται οι σχετικές δημοσιεύσεις, όπως στο Σχήμα 5.2. Ιδιαίτερο χαρακτηριστικό του συστήματος είναι ότι χρησιμοποιεί την απήχηση με διαφορετικούς τρόπους: υποστηρίζεται εμφάνιση των αποτελεσμάτων με βάση τη δημοφιλία ή την επιρροή καθώς και η εφαρμογή φίλτρων στα αποτελέσματα με βάση το επίπεδο δημοφιλίας, ή επιρροής τους. Οι χρήστες μπορούν να διαλέξουν το κριτήριο κατάταξης των αποτελεσμάτων (το είδος απήχησης),

[7] https://radimrehurek.com/gensim/

[8] http://lucene.apache.org/solr/



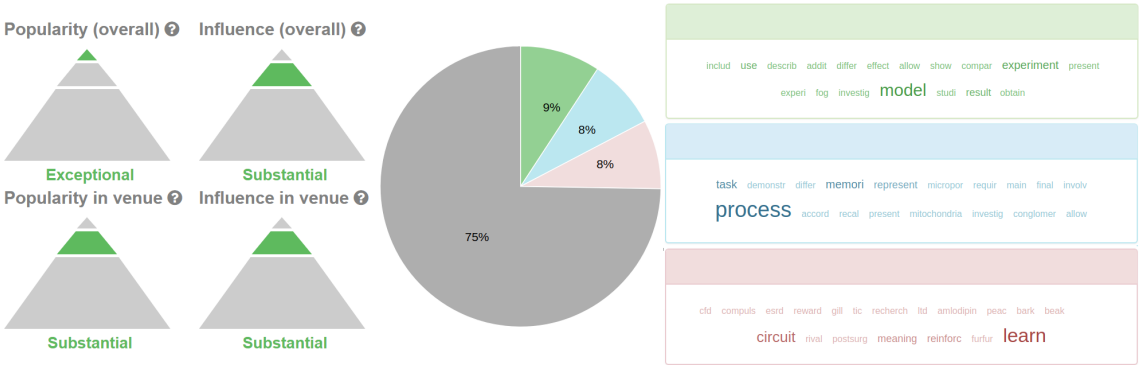

Σχήμα 5.2: Αποτελέσματα αναζήτησης BIP! Finder.

ενώ τους δίνεται ακόμα η επιλογή να το συνδυάζουν με τη σχετικότητα, πατώντας στα κατάλληλα κουμπιά κάτω από το πεδίο αναζήτησης. Οι χρήστες μπορούν ακόμα να αποκλείσουν από τα αποτελέσματα μιας αναζήτησης τις δημοσιεύσεις που έχουν χαμηλές βαθμολογίες απήχησης, χρησιμοποιώντας τα φίλτρα στην αριστερή πλευρά της οθόνης. Τέλος, διατίθενται επιπλέον φίλτρα που δίνουν τη δυνατότητα περιορισμού των αποτελεσμάτων με βάση τα έτη δημοσίευσης, ή με βάση το επιστημονικό περιοδικό.

**Σύγκριση Δημοσιεύσεων.** Οι χρήστες του BIP! μπορούν να επιλέξουν δύο έως τέσσερις δημοσιεύσεις και να τις συγκρίνουν ως προς την απήχηση και άλλα χαρακτηριστικά. Μετά την επιλογή των δημοσιεύσεων, ο χρήστης πατάει στο κουμπί σύγκρισης ("Compare"), το οποίο τον ανακατευθύνει σε νέο παράθυρο. Η σελίδα που εμφανίζεται στο νέο παράθυρο παρουσιάζει κάποια «διαγράμματα ραντάρ» (Radar Charts) τα οποία παρουσιάζουν συγκριτικά κάποια ενδιαφέροντα χαρακτηριστικά των επιλεγμένων δημοσιεύσεων, όπως τις βαθμολογίες δημοφιλίας και επιρροής και τον αριθμό επισκέψεων της σελίδας κάθε μιας στο BIP!. Επιπλέον δίνονται διαγράμματα που παρουσιάζουν συγκριτικά τους αριθμούς αναφορών ανά έτος που έχει λάβει η κάθε δημοσίευση.

**Λεπτομέρειες Δημοσίευσης.** Οι χρήστες του BIP! μπορούν να δουν διάφορες λεπτομέρειες και ενημερωτικά γραφικά (Infographics) για δημοσιεύσεις που επιλέγουν. Πατώντας πάνω στο κάθε αποτέλεσμα αναζήτησης, οι χρήστες ανακατευθύνονται σε μια σελίδα στην οποία μπορούν να δουν την περίληψη της δημοσίευσης, τους συγγραφείς, τη γλώσσα συγγραφής και το περιοδικό όπου δημοσιεύτηκε, καθώς και σειρά γραφημάτων όπως αυτά που παρουσιάζονται στο Σχήμα 5.3. Αυτά περιλαμβάνουν πυραμίδες στις οποίες απεικονίζεται αν η επιλεγμένη δημοσίευση βρίσκεται ανάμεσα στο καλύτερο $0,01\%, 1\%$, ή στο υπόλοιπο $99\%$ των δημοσιεύσεων όσο αφορά τη δημοφιλία και την επιρροή, τόσο στο σύνολο των δημοσιεύσεων, όσο και μεταξύ όλων των δημοσιεύσεων που εκδόθηκαν από το ίδιο περιοδικό. Επιπλέον, παρουσιάζεται το ιστορικό των αναφορών που έχει λάβει η επιλεγμένη δημοσίευση κάθε έτος από την έκδοσή της και μετά. Σ' αυτό το γράφημα έχουν προστεθεί δύο «εικονικές τροχιές αναφορών» που αφορούν μια εικονική δημοσίευση που βρίσκεται στο καλύτερο $0.01\%$ και μια που βρίσκεται στο καλύτερο $1\%$ των δημοσιεύσεων, με βάση την επιρροή. Η προσθήκη αυτή δίνει μια καλύτερη διαίσθηση της δυναμικής της υπό εξέτασης δημοσίευσης. Τέλος, περιλαμβάνεται πληροφορία που έχει προκύψει από την ανάλυση LDA για τη δημοσίευση, η οποία παρέχει στο χρήστη μια εικόνα για τις λέξεις που απαρτίζουν τα σημαντικότερα λανθάνοντα θέματα τα οποία πραγματεύεται η δημοσίευση (Σχήμα 5.3).

**Φάκελοι Αγαπημένων.** Οι χρήστες που έχουν φτιάξει λογαριασμό στο σύστημα, έχουν τη δυνατότητα να οργανώνουν τις δημοσιεύσεις που τους ενδιαφέρουν σε θεμα-




Σχήμα 5.3: Ενημερωτικά γραφήματα για δημοσιεύσεις στο BIP!.

τικούς φακέλους. Μια δημοσίευση εισάγεται στα αγαπημένα, όταν ο χρήστης πατήσει το αντίστοιχο κουμπί που συνοδεύει κάθε αποτέλεσμα αναζήτησης.

**Προγραμματιστική Διεπαφή (API) Βαθμολογιών.** Το σύστημα παρέχει ένα API το οποίο δίνει πρόσβαση σε όλες τις βαθμολογίες PageRank και RAM που έχουν υπολογιστεί για τις δημοσιεύσεις που βρίσκονται στη βάση δεδομένων του BIP!. Το API είναι ελεύθερα προσβάσιμο[9] και αναπτύχθηκε βασισμένο σε αρχιτεκτονική μικροϋπηρεσιών (Microservices), ως ανεξάρτητη εφαρμογή Node.js. Ο σκοπός του είναι να δοθεί κίνητρο για ανάπτυξη επιπλέον εφαρμογών από τρίτους, παρέχοντας στην ερευνητική κοινότητα μια «εργαλειοθήκη» για την ανάπτυξη εφαρμογών ερευνητικής αναλυτικής.

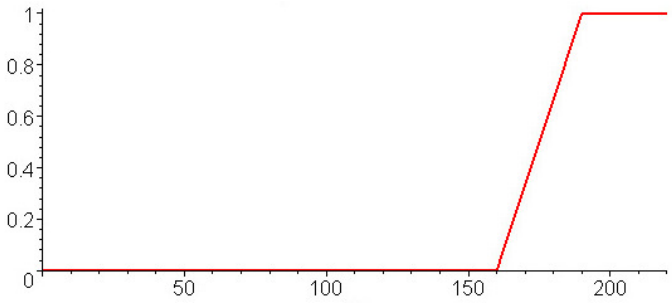
#### 5.1.1.3 Συνδυασμός Απήχησης και Σχετικότητας Με το Ερώτημα

Το σύστημα BIP! δίνει τη δυνατότητα ιεράρχησης αποτελεσμάτων συνδυάζοντας την απήχηση των δημοσιεύσεων με τη σχετικότητά τους με το εκάστοτε ερώτημα. Στην ενότητα αυτή περιγράφουμε το μηχανισμό με βάση τον οποίο γίνεται ο συνδυασμός των βαθμολογιών. Ο μηχανισμός αυτός βασίζεται στη θεωρία των ασαφών συνόλων (Fuzzy Sets), βασικές αρχές της οποίας παρουσιάζουμε συνοπτικά στη συνέχεια.

**Ασαφή Σύνολα.** Τα ασαφή σύνολα αποτελούν μια επέκταση της κλασσικής συνολοθεωρίας, η οποία επιτρέπει την μαθηματική περιγραφή ιδιοτήτων του πραγματικού κόσμου, οι οποίες αν και συνηθίζονται στη καθημερινή συλλογιστική των ανθρώπων, ωστόσο δε μπορούν να περιγραφούν σαφώς μέσω της κλασσικής συνολοθεωρίας. Για παράδειγμα, έννοιες που αντιστοιχούν σε ιδιότητες όπως «ψηλός» δεν περιγράφονται επαρκώς από την κλασσική συνολοθεωρία. Αυτό γιατί η κλασσική συνολοθεωρία θα απαιτούσε να οριστεί κάποιο σαφές όριο ύψους με βάση το οποίο οι άνθρωποι κατηγοριοποιούνται ως μέλη των συνόλων «ψηλός» και «όχι ψηλός». Αντιθέτως τα μέλη των ασαφών συνόλων χαρακτηρίζονται από ένα φάσμα βαθμών συμμετοχής σε αυτά.

Τυπικά ένα ασαφές σύνολο ορίζεται ως εξής: έστω ένα σύνολο αντικειμένων $X$ και $x$ ένα από τα αντικείμενα του $X$. Ασαφές σύνολο είναι μια κλάση $A$ στο $X$ που καθορίζεται από τη συνάρτηση συμμετοχής $f_A(x)$, η οποία αντιστοιχίζει κάθε αντικείμενο του $X$ σε έναν πραγματικό αριθμό, $f_A(x) \in [0, 1]$, που αναπαριστά το βαθμό συμμετοχής του $x$ στο $A$ [91]. Όσο πιο κοντά στη μονάδα είναι η τιμή της συνάρτησης συμμετοχής του $x$ στο $A$, τόσο μεγαλύτερη είναι η συμμετοχή του αντικειμένου στο σύνολο. Στις ειδικές περιπτώσεις όπου η τιμή είναι 0 και 1, τότε έχουμε μη συμμετοχή και πλήρη συμμετοχή του αντικειμένου στο σύνολο, αντίστοιχα. Ένα παράδειγμα

[9] http://bip.imsi.athenarc.gr:4000/documentation



Σχήμα 5.4: Παράδειγμα συνάρτησης συμμετοχής στο ασαφές σύνολο «ψηλός».

συνάρτησης συμμετοχής που περιγράφει την έννοια «ψηλός άνθρωπος» δίνεται στο Σχήμα 5.4. Με βάση τη συνάρτηση συμμετοχής του σχήματος, οι άνθρωποι με ύψος μεγαλύτερο από $1,60$ μέτρα θεωρείται ότι έχουν κάποια συμμετοχή στο σύνολο, η οποία αυξάνει με το ύψος. Άνθρωποι με ύψος μεγαλύτερο από $1,88$ θεωρείται πως έχουν πλήρη συμμετοχή στο σύνολο, ενώ όσοι έχουν ύψος μικρότερο από $1,60$ δεν θεωρούνται ψηλοί.

Σε αντιστοιχία με την κλασσική συνολοθεωρία, ορίζονται ορισμένες ιδιότητες των ασαφών συνόλων (λεπτομέρειες για τις αποδείξεις τους παρουσιάζονται στο [91]). Οι βασικότερες είναι οι ακόλουθες:

- Ένα ασαφές σύνολο είναι κενό αν η συνάρτηση συμμετοχής του δίνει πάντοτε την τιμή μηδέν για όλα τα αντικείμενα του χώρου $X$.
- Δύο ασαφή σύνολα είναι ίδια όταν ταυτίζονται οι συναρτήσεις συμμετοχής τους.
- Το συμπλήρωμα ενός ασαφούς συνόλου ορίζεται από τη συνάρτηση συμμετοχής $f_{A'} = 1 - f_A$.
- Η ένωση δύο ασαφών συνόλων $A$ και $B$ με συναρτήσεις συμμετοχής $f_A, f_B$ ορίζεται από τη συνάρτηση συμμετοχής $f_C = Max(f_A, f_B)$.
- Η τομή δύο ασαφών συνόλων $A$ και $B$ με συναρτήσεις συμμετοχής $f_A, f_B$ ορίζεται από τη συνάρτηση συμμετοχής $f_C = Min(f_A, f_B)$.

Από τις παραπάνω ιδιότητες η τελευταία είναι και αυτή που θα μας απασχολήσει.

**Συνδυασμός Απήχησης - Σχετικότητας.** Όπως αναφέρθηκε νωρίτερα, το BIP! βασίζεται στο Solr για να εντοπίσει δημοσιεύσεις σχετικές με τις λέξεις-κλειδιά που εισάγει ο χρήστης. Η μηχανή Solr βασίζεται στη βιβλιοθήκη ευρετηριοποίησης και αναζήτησης Lucene της Java. Η τελευταία υλοποιεί παραδοσιακούς αλγορίθμους από την ανάκτηση πληροφορίας για να βαθμολογήσει τη σχετικότητα των αποτελεσμάτων με το ερώτημα αναζήτησης. Οι βαθμολογίες αυτές συνήθως προκύπτουν από κάποιον τύπο που βασίζεται στο TF - IDF, δηλαδή συνδυάζει τη συχνότητα των όρων που αναζητήθηκαν στο εκάστοτε κείμενο, με τη συχνότητα εμφάνισης των όρων γενικά στο σύνολο των κειμένων. Όταν ο χρήστης του BIP! πραγματοποιεί μια αναζήτηση επομένως, αρχικά χρησιμοποιείται το Solr για να επιστραφεί το σύνολο των σχετικών δημοσιεύσεων και στη συνέχεια αυτές συνδυάζονται επιπλέον με τη βαθμολογία απήχησής τους, που αποθηκεύεται στη σχεσιακή βάση του συστήματος.

Τόσο οι βαθμολογίες σχετικότητας, όσο και οι βαθμολογίες απήχησης, μπορούν εύκολα να μετατραπούν σε τιμές μιας συνάρτησης συμμετοχής ενός ασαφούς συνόλου: αφού τις κανονικοποιήσουμε (διαιρέσουμε με τις αντίστοιχες μέγιστες τιμές για όλα

τα αποτελέσματα), προκύπτει ένα σύνολο βαθμολογιών στο διάστημα $[0, 1]$. Αυτό μας επιτρέπει να θεωρήσουμε ότι οι τιμές αυτές αποτελούν τις πραγματικές τιμές της συνάρτησης συμμετοχής δύο ασαφών συνόλων: του συνόλου που περιγράφει δημοσιεύσεις «σχετικές» και του συνόλου που περιγράφει δημοσιεύσεις «μεγάλης απήχησης».[10] Κάθε δημοσίευση έχει βαθμό συμμετοχής σε κάθε ένα από τα ασαφή σύνολα, ίσο με την τιμή της αντίστοιχης βαθμολογίας.

Η απαίτηση του χρήστη που θέλει τα αποτελέσματα να συνδυάζουν τη σχετικότητα και την απήχηση, καλύπτεται όταν ιεραρχούνται ψηλότερα δημοσιεύσεις που είναι σχετικές *και ταυτόχρονα* έχουν μεγάλη απήχηση. Αν εξετάσουμε αυτές τις δύο ιδιότητες μέσω της θεωρίας των ασαφών συνόλων, τότε ο χρήστης ενδιαφέρεται να δει πρώτες τις δημοσιεύσεις με το μεγαλύτερο βαθμό συμμετοχής στην τομή τους. Επομένως, με δεδομένο ένα σύνολο δημοσιεύσεων $U$, που αποτελούν τα αποτελέσματα μιας αναζήτησης, θεωρούμε ότι κάθε δημοσίευση $p_i \in U$ έχει μια βαθμολογία σχετικότητας με το ερώτημα $rel(p_i)$ και μια βαθμολογία απήχησης $imp(p_i)$ (οι οποίες αντιστοιχούν στους βαθμούς συμμετοχής στα αντίστοιχα ασαφή σύνολα). Θα μπορούσαμε να καθορίσουμε την ιεράρχηση του συνόλου $U$ χρησιμοποιώντας τον βαθμό συμμετοχής στην τομή των ασαφών συνόλων, η οποία έχει συνάρτηση συμμετοχής:

$$f_U = min(rel(p_i), imp(p_i)) \tag{5.1}$$

Στην πράξη με αυτό τον τρόπο οι βαθμολογίες σχετικότητας επηρεάζουν ελάχιστα το αποτέλεσμα. Αυτό συμβαίνει διότι, όπως αναφέραμε και στο Κεφάλαιο 4 οι βαθμολογίες απήχησης ακολουθούν μια κατανομή εκθετικού νόμου που χαρακτηρίζεται από τη «μεγάλη ουρά». Συνέπεια αυτού είναι οι περισσότερες δημοσιεύσεις να έχουν πολύ μικρή βαθμολογία απήχησης και έτσι να κυριαρχεί η συγκεκριμένη βαθμολογία στον καθορισμό του ελάχιστου μεταξύ $rel(p_i)$ και $imp(p_i)$. Για το λόγο αυτό αξιοποιούμε το γεγονός ότι οι κανονικοποιημένες βαθμολογίες απήχησης έχουν πεδίο τιμών το $[0, 1]$. Παίρνουμε στη θέση της βαθμολογίας $imp(p_i)$ τη βαθμολογία $imp^{\frac{1}{4}}(p_i)$ και ορίζουμε την χαρακτηριστική συνάρτηση της τομής των ασαφών συνόλων «σχετικές δημοσιεύσεις» και «δημοσιεύσεις υψηλής απήχησης» ως:

$$f_U = min(rel(p_i), imp^{\frac{1}{4}}(p_i)) \tag{5.2}$$

Η χαρακτηριστική συνάρτηση που περιγράφεται από την Εξίσωση 5.2 δίνει τους βαθμούς συμμετοχής των αποτελεσμάτων αναζήτησης στο σύνολο «σχετικές δημοσιεύσεις μεγάλης απήχησης», ο οποίος αποτελεί και τη βάση με την οποία το BIP! ιεραρχεί αποτελέσματα. Σε όλα τα παραπάνω, ως $imp(p_i)$ βαθμολογία χρησιμοποιείται, ανάλογα με την επιλογή του χρήστη, η (κανονικοποιημένη) βαθμολογία PageRank ή RAM.

### 5.1.2 mirPub v2

#### 5.1.2.1 Εισαγωγή: miRNAs και εξέλιξη δεδομένων

Τα βιομόρια miRNA είναι μικρά μόρια RNA στα κύτταρα, τα οποία παρεμποδίζουν την έκφραση γονιδίων [12] και ως εκ τούτου σχετίζονται με την εμφάνιση διαφόρων ασθενειών, όπως διάφορα είδη καρκίνου [94]. Εξαιτίας της βιολογικής σημασίας τους, πολλές βάσεις δεδομένων συλλέγουν πληροφορία σχετικά με αυτά [45, 76, 83]. Καθώς

[10] Κάνουμε τη θεώρηση ότι όλες οι δημοσιεύσεις συμμετέχουν σε κάποιο βαθμό και στα δύο ασαφή σύνολα.

τα miRNA αποτελούν μια σχετικά πρόσφατη[11] ανακάλυψη, η επιστημονική γνώση γύρω από αυτά διαρκώς εμπλουτίζεται. Κατ΄ αυτόν τον τρόπο εισάγονται και νέοι κανόνες ονοματολογίας για τα miRNAs (οι οποίοι μπορεί να αντανακλούν κάποια νέα πληροφορία βιολογικής φύσης). Το αποτέλεσμα αυτής της κατάστασης είναι ότι ένας ερευνητής που δεν είναι πάντοτε ενήμερος για τους τρέχοντες κανόνες ονοματολογίας, μπορεί να αναφέρεται σε κάποιο miRNA με ένα παρωχημένο όνομα, ή ανάστροφα, γνωρίζοντας το τρέχον όνομα να αγνοεί παλιότερα που αναφέρονται στο ίδιο miRNA. Έτσι, όταν αναζητά βιβλιογραφία για ένα miRNA το οποίο μελετάει, ενδέχεται να αγνοεί βιβλιογραφία που αναφέρεται σε αυτό με παλιότερα ή πιο πρόσφατα ονόματα από αυτά που γνωρίζει.

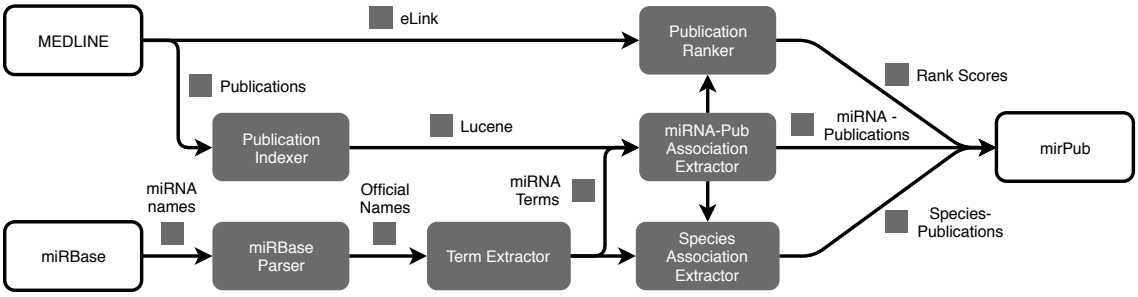
#### 5.1.2.2 mirPub

Το mirPub [75] είναι μια μηχανή βιβλιογραφικής αναζήτησης για miRNAs, η οποία στοχεύει να παρέχει στους χρήστες τη πληρέστερη δυνατή βιβλιογραφία που σχετίζεται με τα miRNAs τα οποία μελετάνε. Για να το επιτύχει αυτό, το σύστημα πραγματοποιεί αυτόματη επέκταση των όρων αναζήτησης, λαμβάνοντας υπόψη την εξέλιξη των ονομάτων που καταγράφονται για τα miRNAs. Τα βασικά κομμάτια λογισμικού από τα οποία αποτελείται το BIP! και στα οποία στηρίζεται η λειτουργία του, είναι: το σύστημα επέκτασης ονομάτων και το σύστημα αντιστοίχησης όρων (ονομάτων miRNA) με επιστημονικές δημοσιεύσεις. Το σύστημα επέκτασης ονομάτων βασίζεται στην επεξεργασία των αρχείων των διαφορετικών εκδόσεων της κύριας βάσης δεδομένων για miRNAs, της mirBase [30]. Το σύστημα αντιστοίχισης όρων με δημοσιεύσεις βασίζεται σε δεδομένα της MEDLINE (του αμερικάνικου οργανισμού υγείας[12], η οποία καταγράφει επιστημονικές δημοσιεύσεις των επιστημών ζωής) και την βιβλιοθήκη κατασκευής ευρετηρίων Lucene. Το σύστημα συνολικά καταγράφει $20,690$ επιστημονικές δημοσιεύσεις, οι οποίες συσχετίζονται με $31,984$ όρους-ονόματα miRNA.

Η επέκταση της αρχικής εφαρμογής, που ονομάστηκε mirPub v2, προσθέτει δύο επιπλέον τμήματα λογισμικού: το σύστημα συσχέτισης κειμένων με οργανισμούς (Species Association Extractor) και το σύστημα κατάταξης δημοσιεύσεων (Publication Ranker). Το πρώτο σύστημα σχετίζεται με το γεγονός ότι πολλά ονόματα miRNA που εμφανίζονται σε δημοσιεύσεις δε κωδικοποιούν πληροφορία για τον οργανισμό αναφοράς του βιομορίου. Το λογισμικό εξαγωγής συσχετίσεων οργανισμών με δημοσιεύσεις πραγματοποιεί εξόρυξη κειμένου (Text Mining) για να εντοπίσει σύνολα συνωνύμων που αναφέρονται σε διάφορους οργανισμούς, στις δημοσιεύσεις που αναλύει. Η καταγραφή των συσχετίσεων δημοσιεύσεων-οργανισμών δίνει τη δυνατότητα εφαρμογής φίλτρων που διευκολύνουν τους ερευνητές στην αναζήτηση βιβλιογραφίας για miRNA σε συγκεκριμένους οργανισμούς. Το σύστημα κατάταξης δημοσιεύσεων είναι υπεύθυνο για τη συλλογή αναφορών μεταξύ των δημοσιεύσεων που καταγράφονται στο σύστημα και την παραγωγή βαθμολογιών κατάταξης με διάφορες μεθόδους (περισσότερες λεπτομέρειες στην Ενότητα 5.1.2.3). Με το λογισμικό αυτό εξασφαλίζεται ότι σε αναζητήσεις, όπου (λόγω της επέκτασης ονομάτων) έχουμε χιλιάδες δημοσιεύσεις ως αποτελέσματα που σχετίζονται με μια λέξη-κλειδί, οι χρήστες θα εντοπίσουν γρήγορα τις σημαντικότερες ανάμεσά τους. Μια συνοπτική περιγραφή της αρχιτεκτονικής του mirPub δίνεται στο Σχήμα 5.5.

[11] Το πρώτο miRNA ανακαλύφθηκε τη δεκαετία του ΄90 ενώ για πρώτη φορά αναγνωρίστηκε ο βιολογικός τους ρόλος μια δεκαετία περίπου αργότερα.

[12] http://www.ncbi.nlm.nih.gov/pubmed




Σχήμα 5.5: Αρχιτεκτονική και Λογισμικά της Μηχανής Αναζήτησης mirPub

#### 5.1.2.3 Δίκτυο Αναφορών, Μέθοδοι και Παραμετροποίηση

Το mirPub παρέχει τη δυνατότητα στους χρήστες να επιλέγουν με βάση ποια μέθοδο κατάταξης θα εμφανίζονται τα αποτελέσματα αναζήτησης. Οι βαθμολογίες που αποδίδει στις δημοσιεύσεις η κάθε μέθοδος κατάταξης υπολογίστηκαν μετά από ανάλυση του δικτύου αναφορών που αποτελείται από τις δημοσιεύσεις που καταγράφει το mirPub και τις μεταξύ τους αναφορές. Η κατασκευή αυτού του δικτύου αναφορών έγινε με χρήση της προγραμματιστικής διεπαφής εφαρμογών (Application Programming Interface - API) eLink[13] που παρέχει η MEDLINE (Σχήμα 5.5). Σημειώνουμε ότι ο γράφος αυτός αποτελεί έναν υπογράφο (αραιότερο από πλευράς κόμβων και ακμών) σε σχέση με τον πραγματικό, πλήρη γράφο δημοσιεύσεων της MEDLINE, καθώς περιέχει μόνο αναφορές από και προς δημοσιεύσεις που καταγράφει το mirPub.

Το σύστημα ενσωματώνει δυνατότητες κατάταξης αποτελεσμάτων με βάση τις ακόλουθες μεθόδους: την κλασσική PageRank (Εξίσωση 2.1) και δύο τροποποιήσεις της μεθόδου YetRank [37] (Ενότητες 3.1.2 και 3.1.3). Συγκεκριμένα, οι παραλλαγές της μεθόδου YetRank βασίστηκαν στους ακόλουθους τύπους:

$$s_i = \alpha \sum_j P_{i,j} s_j + (1-\alpha) \cdot IF_i \cdot e^{-\frac{t-t_i}{\tau}} \tag{5.3}$$

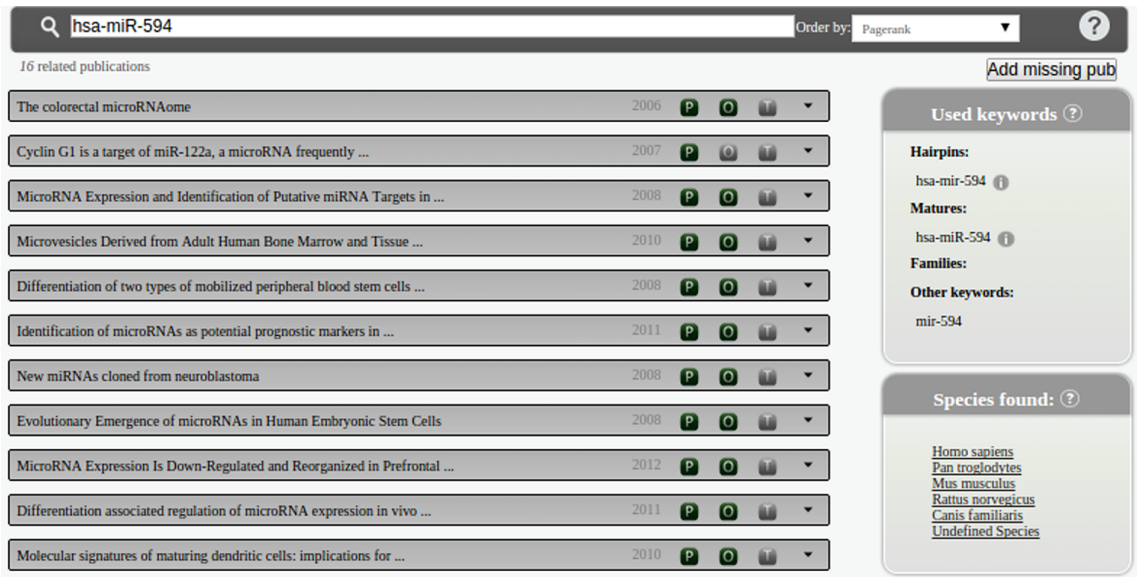
και

$$s_i = \alpha \sum_p P_{i,j} s_j + \beta \cdot IF_i + \gamma \cdot e^{-\frac{t-t_i}{\tau}} \tag{5.4}$$

όπου $IF_i$ είναι η τιμή του Impact Factor του περιοδικού της δημοσίευσης $i$ και $\tau$ είναι ένας παράγοντας γήρανσης. Στην Εξίσωση 5.4 θέτουμε $\alpha + \beta + \gamma = 1$.

Για να ορίσουμε τις τιμές των μεταβλητών παραμέτρων των εξισώσεων 5.3 και 5.4 στηριχθήκαμε στο πλαίσιο αξιολόγησης που ορίσαμε στο Κεφάλαιο 4. Συγκεκριμένα, βασιστήκαμε στη χρήση του PageRank ως κατάλληλης κεντρικότητας που αποτυπώνει την επιρροή. Χρησιμοποιήσαμε τη μέθοδο παρακράτησης δεδομένων (Ενότητα 3.2) τρέχοντας τις παραλλαγές του YetRank στο δίκτυο αναφορών που περιέχει όλες τις δημοσιεύσεις, εκτός από αυτές των τελευταίων πέντε ετών. Στη διαδικασία αυτή έγινε δοκιμή διαφόρων παραμετροποιήσεων. Από αυτές επιλέξαμε αυτή που δίνει την βέλτιστη συσχέτιση (Spearman's $\rho$) με την κατάταξη βάσει του PageRank στο σύνολο του δικτύου αναφορών (υπόβαθρο αληθείας επιρροής).

[13] https://www.ncbi.nlm.nih.gov/books/NBK25499/#_chapter4_ELink_




Σχήμα 5.6: Σελίδα αποτελεσμάτων του mirPub v2 για αναζήτηση με τον όρο "hsa-miR-594".

#### 5.1.2.4 Διεπαφή Χρήστη

Στο Σχήμα 5.6 παρουσιάζουμε τη διεπαφή χρήστη του mirPub. Συγκεκριμένα, παρουσιάζονται τα αποτελέσματα αναζήτησης για τον όρο "hsa-miR-594". Ο χρήστης μπορεί να ταξινομήσει τα αποτελέσματα της αναζήτησης είτε χρονολογικά, είτε χρησιμοποιώντας μια από τις μεθόδους κατάταξης που υλοποιήσαμε, διαλέγοντας την αντίστοιχη επιλογή από το μενού που βρίσκεται στα δεξιά του πεδίου αναζήτησης. Επιπλέον, ο χρήστης μπορεί να επιλέξει να εμφανίζονται μόνο οι δημοσιεύσεις που αναφέρονται στο miRNA σε συγκεκριμένο οργανισμό, τον οποίο μελετάει, διαλέγοντας την κατάλληλη επιλογή από το μενού στα δεξιά των αποτελεσμάτων αναζήτησης.

## 5.2 Μέθοδοι Αποτελεσματικής Κατάταξης με Βάση τη Δημοφιλία

Σε αυτή την ενότητα επικεντρωνόμαστε στο πρόβλημα της κατάταξης επιστημονικών δημοσιεύσεων με βάση τη δημοφιλία. Η ανάλυση που προηγήθηκε στο Κεφάλαιο 4 ανέδειξε ένα αισθητό περιθώριο βελτίωσης στην αποτελεσματικότητα παραγωγής κατατάξεων με βάση τη δημοφιλία σε σχέση με τις τρέχουσες μεθόδους στη βιβλιογραφία. Στο υπόλοιπο του παρόντος κεφαλαίου εστιάζουμε σε αυτό το πρόβλημα και παρουσιάζουμε μια νέα, αποτελεσματικότερη μέθοδο για κατάταξης με βάση τη δημοφιλία.

Όπως είδαμε στο Κεφάλαιο 3, για την αντιμετώπιση της μεροληψίας των μεθόδων κατάταξης υπέρ των πιο παλιών δημοσιεύσεων, η οποία δυσχεραίνει την παραγωγή κατατάξεων με βάση τη δημοφιλία, συνήθως εισάγονται χρονικοί παράγοντες σε τροποποιήσεις κλασσικών μεθόδων, όπως στον αριθμό αναφορών, ή στο PageRank. Ειδικά στη περίπτωση του PageRank η εισαγωγή χρονικών παραγόντων αντανακλά κάποιες παραδοχές που γίνονται σε σχέση με τη συμπεριφορά ενός «τυχαίου ερευνητή». Για παράδειγμα μέθοδοι όπως οι CiteRank και FutureRank κάνουν τη θεώρηση ότι ο ερευνητής όταν δεν επιλέγει να διαβάσει κάποια δημοσίευση από μια λίστα αναφορών, επιλέγει κάποια από όλο το δίκτυο αναφορών, με προτίμηση όμως στις πιο καινούριες.

Πίνακας 5.1: Αριθμός των δημοφιλών δημοσιεύσεων (βάσει του υπόβαθρου αληθείας P-CC) οι οποίες βρίσκονται ανάμεσα στις 100 πρώτες σε αριθμό αναφορών τα τελευταία 5 έτη.

| **Σύνολο Δεδομεων** | **hep-th** | **APS** | **PMC** | **DBLP** |
|---|---|---|---|---|
| **Τελευταία Δημοφιλείς** | 41 | 54 | 54 | 63 |

Εδώ υποστηρίζουμε ότι υπάρχει ένας επιπλέον, ανεκμετάλλευτος, μηχανισμός που επηρεάζει τη συσσώρευση αναφορών. Συγκεκριμένα θεωρούμε ότι σε μεγάλο βαθμό η κατανομή των αναφορών μεταξύ των δημοσιεύσεων στα πιο πρόσφατα έτη καθορίζει σε μεγάλο βαθμό και τη συσσώρευση αναφορών στο άμεσο μέλλον. Με άλλα λόγια, ο βαθμός στον οποίο οι δημοσιεύσεις τραβάνε την *προσοχή της ερευνητικής κοινότητας (στο εξής απλά «προσοχή»)* στη τρέχουσα χρονική περίοδο, δεν αναμένεται να μεταβληθεί σημαντικά στο πολύ κοντινό μέλλον. Εξερευνούμε στη συνέχεια αυτή την υπόθεση και βρίσκουμε ότι ισχύει σε ένα σημαντικό βαθμό σε διάφορους γράφους αναφορών. Εν συνεχεία, εισάγουμε ένα μηχανισμό βασισμένο στην προσοχή (Attention-based Mechanism) στο μοντέλο του PageRank, επιπρόσθετα με τους χρονικούς παράγοντες. Ο μηχανισμός αυτός μοιάζει με μια χρονικά περιορισμένη εκδοχή του λεγόμενου μηχανισμού της προτιμησιακής προσκόλλησης (Preferential Attachment) [4] και προσομοιώνει το γεγονός ότι δημοσιεύσεις που λαμβάνουν στο τελευταίο χρονικό διάστημα αναφορές συνεχίζουν να λαμβάνουν αναφορές και στο κοντινό μέλλον.

Η μέθοδος που προτείνουμε αποτελεί τροποποίηση του PageRank και προσομοιώνει έναν ερευνητή, ο οποίος ξεκινάει να διαβάζει μια δημοσίευση και στη συνέχεια φέρεται με έναν από τους ακόλουθους τρεις τρόπους: είτε επιλέγει να διαβάσει μια άλλη δημοσίευση από τη λίστα αναφορών της πρώτης, είτε επιλέγει μια οποιαδήποτε δημοσίευση με προτίμηση στις πιο πρόσφατες, είτε επιλέγει μια οποιαδήποτε δημοσίευση με προτίμηση σε αυτές που έλαβαν πολλές αναφορές πρόσφατα. Η μέθοδος μας αξιολογείται ως προς την αποτελεσματικότητά της στην Ενότητα 5.2.4, κατά τα πρότυπα του Κεφαλαίου 4.

### 5.2.1 Το Διάνυσμα Πρόσφατου Ενδιαφέροντος (Διάνυσμα Προσοχής)

Βασική υπόθεση εργασίας μας στην τρέχουσα ενότητα είναι ότι οι ερευνητές έχουν την τάση να διαβάζουν και να αναφέρουν εργασίες που τον τελευταίο καιρό τραβάνε σε σημαντικό βαθμό την προσοχή της επιστημονικής κοινότητας. Για να διερευνήσουμε σε τι βαθμό ισχύει η υπόθεση αυτή εξετάζουμε τα τέσσερα σύνολα δεδομένων που χρησιμοποιήσαμε στην πειραματική αξιολόγηση του Κεφαλαίου 4. Χωρίζουμε τα σύνολα δεδομένων έτσι ώστε ο λόγος $\eta$ (μελλοντική/τρέχουσα κατάσταση, Κεφάλαιο 4) να παίρνει την τιμή βάσης του $\eta = 1.6$. Στη συνέχεια εξετάζουμε την τομή του συνόλου των 100 δημοσιεύσεων με τις περισσότερες αναφορές στα τελευταία πέντε έτη, με αυτό των 100 πρώτων δημοσιεύσεων με βάση το υπόβαθρο αληθείας P-CC για τη δημοφιλία. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 5.1.

Από τον πίνακα παρατηρούμε ότι σε κάθε σύνολο δεδομένων η τομή που εξετάζουμε είναι αρκετά μεγάλη. Η παρατήρηση αυτή επιβεβαιώνει την υπόθεσή μας ότι ο βαθμός στον οποίο μια δημοσίευση τραβάει την προσοχή της επιστημονικής κοινότητας στο παρόν είναι ενδεικτικός του ενδιαφέροντος (και άρα των αναφορών) που θα τραβήξει στο βραχυχρόνιο μέλλον. Για να εκμεταλλευτούμε το γεγονός αυτό πρέπει να πο-

σοτικοποιήσουμε την έννοια της προσοχής. Ορίζουμε την προσοχή ως εξής: έστω ότι $\mathbf{A}[t_N - y : t_N]$ είναι ο πίνακας γειτνίασης για ένα δίκτυο αναφορών, ο οποίος λαμβάνει υπόψη μόνο τις αναφορές που έγιναν στα τελευταία $y$ χρόνια. Η *προσοχή* μιας δημοσίευσης $p_i$ υπολογίζεται ως:

$$\Pi(p_i) = \frac{\sum_j A_{i,j}[t_N - y : t_N]}{\sum_i \sum_j A_{i,j}[t_N - y : t_N]}, \tag{5.5}$$

Η Εξίσωση 5.5 αντιστοιχεί στο ποσοστό των αναφορών που έλαβε η δημοσίευση $p_i$, επί του συνόλου των αναφορών που έγιναν τα τελευταία $y$ χρόνια προς όλες τις δημοσιεύσεις. Η παράμετρος $y$ εισάγει ένα χρονικό παράγοντα που επιτρέπει, ανάλογα με τη τιμή που θέτουμε, να αποτυπώνουμε τις πιο πρόσφατες τάσεις που παρατηρούνται στο δίκτυο αναφορών. Η παράμετρος αυτή θα πρέπει να ρυθμίζεται κατάλληλα στο εκάστοτε πρόβλημα κατάταξης.

### 5.2.2 Η Μέθοδος Μας

Η μέθοδος μας συνδυάζει τους τρεις μηχανισμούς βάσει των οποίων θεωρείται ότι δρα ένας τυχαίος ερευνητής: ξεκινάει διαβάζοντας μια δημοσίευση, όπως στην περίπτωση του κλασσικού PageRank. Στη συνέχεια μπορεί να επιλέξει (α) να διαβάσει μια άλλη δημοσίευση από τη λίστα αναφορών αυτής που διαβάζει, (β) να επιλέξει μια οποιαδήποτε δημοσίευση του δικτύου, ή (γ) να επιλέξει μια οποιαδήποτε δημοσίευση του δικτύου αναφορών με προτίμηση στις πιο καινούριες. Η δεύτερη περίπτωση ποσοτικοποιείται με το διάνυσμα του πρόσφατου ενδιαφέροντος της εξίσωσης 5.5. Η τελευταία περίπτωση μοντελοποιείται με μια βαθμολογία που εξαρτάται από την ηλικία της κάθε δημοσίευσης, όπως γίνεται στην περίπτωση των μεθόδων CiteRank και FutureRank (Κεφάλαιο 3) και συγκεκριμένα μέσω μιας εκθετικά φθίνουσας συνάρτησης όπως η ακόλουθη:

$$T(p_i) = c \cdot e^{w \cdot (t_N - t_{p_i})}, \tag{5.6}$$

όπου $t_N$ είναι το τρέχον έτος, $t_{p_i}$ το έτος έκδοσης της δημοσίευσης $p_i$, η παράμετρος $w$ είναι ένας αρνητικός πραγματικός αριθμός (καθώς $t_N - t_{p_i} \geq 0$) και η παράμετρος $c$ είναι μια σταθερά κανονικοποίησης έτσι ώστε $\sum_i T(p_i) = 1$. Για να υπολογίσουμε την τιμή που θέτουμε στο $w$ ακολουθούμε μια διαδικασία παρόμοια με αυτή που χρησιμοποιείται στο [66] (περισσότερα στην ενότητα 5.2.4).

Συνδυάζοντας όλους τους παραπάνω μηχανισμούς, θεωρούμε ότι ένας ερευνητής επιλέγει να διαβάσει μια δημοσίευση για έναν από τους παρακάτω λόγους: η δημοσίευση είτε υπήρξε στο επίκεντρο του ενδιαφέροντος πρόσφατα, είτε εκδόθηκε πρόσφατα, είτε βρέθηκε στη λίστα αναφορών άλλης δημοσίευσης που διάβαζε ο ερευνητής. Μοντελοποιούμε το παραπάνω με την ακόλουθη τυχαία διαδικασία: ο ερευνητής, αφού διαβάσει τη δημοσίευση $p_i$ επιλέγει να διαβάσει μια οποιαδήποτε άλλη δημοσίευση από τη λίστα αναφορών της $p_i$ με πιθανότητα $\alpha$. Με πιθανότητα $\beta$, επιλέγει μια δημοσίευση, με βάση την προσοχή που η τελευταία έχει τραβήξει πρόσφατα. Η συμπεριφορά αυτή είναι που κάνει τις δημοσιεύσεις με πολλές πρόσφατες αναφορές να αποκτούν ακόμα περισσότερες, κατά τρόπο παρόμοιο με την προτιμησιακή προσκόλληση (Preferential Attachment) του μοντέλου ανάπτυξης δικτύων που προτάθηκε από τους Barabási-Albert [4]. Τέλος, ο τυχαίος ερευνητής με πιθανότητα $\gamma$ επιλέγει να διαβάσει μια δημοσίευση με προτίμηση στις πιο πρόσφατες.

Το μοντέλο που προτείνουμε υπολογίζει μια βαθμολογία, στην οποία θα αναφερόμαστε ως $AR(p_i)$, για κάθε δημοσίευση $p_i$, με βάση τον ακόλουθο επαναληπτικό υπολογισμό:

$$AR(p_i) = \alpha \cdot \left( \sum_j P_{i,j} \cdot AR(p_j) \right) + \beta \cdot \Pi(p_i) + \gamma \cdot T(p_i), \tag{5.7}$$

όπου $\alpha, \beta, \gamma \in [0, 1]$ και $\alpha + \beta + \gamma = 1$. Όπως γίνεται και με τις μεθόδους που περιγράψαμε στο Κεφάλαιο 3, οι συντελεστές αυτοί ρυθμίζονται μετά από κάποια πειραματική διαδικασία. Επιπλέον στην Εξίσωση 5.7, ο πίνακας $\mathbf{P}$ είναι ο κλασσικός στοχαστικός πίνακας που ορίζεται και στη μέθοδο PageRank.

Αξίζει σε αυτό το σημείο να εξετάσουμε δύο ειδικές τιμές για τον συντελεστή $\beta$. Όταν $\beta = 0$, μια ρύθμιση που ονομάζουμε NO-ATT, το μοντέλο μετατρέπεται σε μια παραλλαγή των μεθόδων που χρησιμοποιούν χρονικούς παράγοντες (Ενότητες 3.1.2-3.1.2.2). Επιπλέον, αν σε αυτή τη περίπτωση θέσουμε και $w = 0$, τότε η μέθοδος «εκφυλίζεται» στη κλασσική μέθοδο PageRank. Η δεύτερη ειδική τιμή που εξετάζουμε, όταν $\beta = 1$, είναι μια ρύθμιση που ονομάζουμε ATT-ONLY. Σε αυτή τη περίπτωση η μέθοδος μας βασίζεται μόνο στο διάνυσμα της πρόσφατης προσοχής, υποθέτοντας ότι οι τρέχουσες τάσεις αναφορών διατηρούνται πλήρως στο μέλλον. Η περίπτωση του ATT-ONLY δεν έχουμε δει να εξετάζεται στην τρέχουσα βιβλιογραφία. Όπως δείχνουμε στην Ενότητα 5.2.4, ο μηχανισμός της προσοχής ακόμα και μόνος του είναι αρκετά ισχυρός, δίνοντας πολλές φορές καλύτερες κατατάξεις ως προς τη δημοφιλία σε σχέση με τις τρέχουσες τεχνολογίες αιχμής. Ωστόσο η περίπτωση όπου $\beta = 1$ δεν είναι ποτέ η βέλτιστη για τη μέθοδο μας, καθώς πάντοτε έχουμε καλύτερα αποτελέσματα όταν συνδυάζουμε το μηχανισμό της προσοχής με τους άλλους δύο που περιγράψαμε παραπάνω.

### 5.2.3 Η Σύγκλιση της Μεθόδου Μας

Η Εξίσωση 5.7 περιγράφει μια επαναληπτική διαδικασία για τον υπολογισμό του διανύσματος $AR$: ξεκινώντας με μια τυχαία τιμή, σε κάθε επανάληψη ενημερώνουμε το διάνυσμα με βάση τη δεξιά πλευρά της Εξίσωσης 5.7. Αυτή η διαδικασία επαναλαμβάνεται μέχρι οι τιμές του διανύσματος $AR$ να συγκλίνουν. Το θεώρημα που ακολουθεί εξασφαλίζει ότι η παραπάνω διαδικασία πάντοτε συγκλίνει.

**Θεώρημα 1.** *Η επαναληπτική διαδικασία που ορίζεται από την Εξίσωση 5.7 συγκλίνει.*

*Απόδειξη.* Μπορούμε να γράψουμε την Εξίσωση 5.7 σε μορφή πολλαπλασιασμού διανύσματος με πίνακα ως εξής:

$$AR = \mathbf{R}\, AR \tag{5.8}$$

όπου ο πίνακας $\mathbf{R}$ ορίζεται ως:

$$R_{i,j} = \alpha \cdot S_{i,j} + \beta \cdot A(p_i) + \gamma \cdot T(p_i) \tag{5.9}$$

Με άλλα λόγια, ο πίνακας $\mathbf{R}$ είναι ένας τροποποιημένος πίνακας μεταβάσεων, τεχνητά εκτεταμένος με κατευθυνόμενες ακμές από κάθε κόμβο προς κάθε άλλον στο δίκτυο. Για κάθε στήλη $c$ του πίνακα $\mathbf{R}$ ισχύει η ακόλουθη ιδιότητα:

$$\begin{aligned}\sum_i R[i,c] &= \alpha \cdot \sum_i S[i,c] + \beta \cdot \sum_i \Pi(p_i) + \gamma \cdot \sum_i T(p_i) \\ &= \alpha + \beta + \gamma = 1\end{aligned} \tag{5.10}$$

Η Εξίσωση 5.10 ισχύει διότι (α) το άθροισμα όλων των βαθμολογιών $\Pi(p_i)$ ισούται εξ' ορισμού με 1, καθώς οι βαθμολογίες αυτές αποτελούν ποσοστά επί ενός συνόλου αναφορών, (β) εξ' ορισμού το άθροισμα των βαθμολογιών $T$ ισούται με 1 και (γ) ο πίνακας $\mathbf{S}$ είναι στοχαστικός ως προς τις στήλες του και άρα επίσης εξ' ορισμού το άθροισμα κάθε στήλης του ισούται με 1. Συνεπώς, ο πίνακας $\mathbf{R}$ επίσης είναι στοχαστικός ως προς τις στήλες του και επομένως η Εξίσωση 5.8 περιγράφει την εφαρμογή της δυναμομεθόδου (Power Method) εφαρμοσμένης στο στοχαστικό πίνακα $\mathbf{R}$.

Η δυναμομέθοδος εφαρμοσμένη σε έναν στοχαστικό πίνακα συγκλίνει υπό τις ακόλουθες συνθήκες [47]: ο πίνακας πρέπει να είναι μη αναγώγιμος και απεριοδικός. Η μη αναγωγιμότητα εξασφαλίζεται για κάθε ισχυρά συνδεδεμένο (Strongly Connected) γράφο, κάτι που ισχύει για το γράφο που περιγράφει ο πίνακας $\mathbf{R}$, καθώς κάθε κόμβος του συνδέεται με όλους τους άλλους. Αυτό συμβαίνει στη περίπτωσή μας λόγω της μοντελοποίησης των τυχαίων μεταβάσεων από κάθε κόμβο προς οποιονδήποτε άλλο κόμβο. Η απεριοδικότητα εξασφαλίζεται για όλους τους γράφους για τους οποίους ισχύει $R[i,i] > 0$. Η συνθήκη αυτή επίσης καλύπτεται από τον πίνακα $\mathbf{R}$, δεδομένου ότι υπάρχουν τεχνητές ακμές από κάθε κόμβο προς όλους τους άλλους.[14] Συνεπώς, ή επαναληπτική διαδικασία εγγυημένα συγκλίνει σε ένα μοναδικό διάνυσμα βαθμολογιών. □

### 5.2.4 Πειραματική Αξιολόγηση

Σε αυτή την ενότητα παρουσιάζουμε μια εξαντλητική πειραματική αξιολόγηση της μεθόδου μας όσο αφορά την αποτελεσματικότητά της να κατατάσσει δημοσιεύσεις με βάση τη δημοφιλία τους. Συγκεκριμένα, στις ακόλουθες υποενότητες:

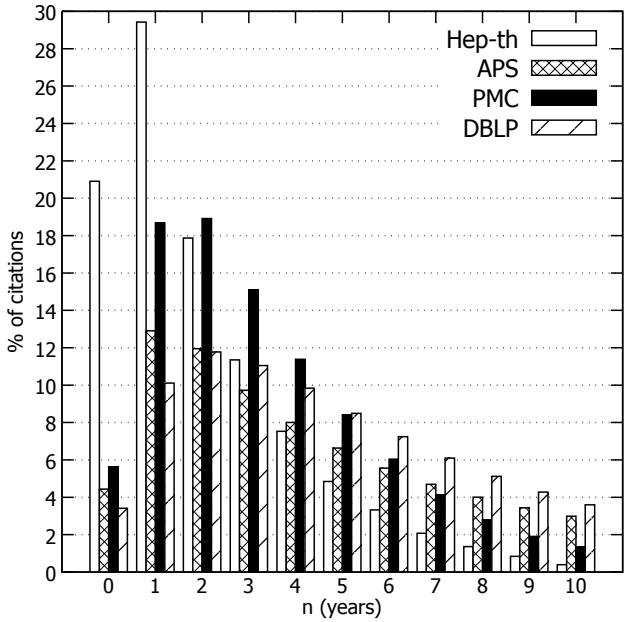
- Περιγράφουμε τη διαδικασία ρύθμισης των παραμέτρων της μεθόδου μας.
- Αξιολογούμε την αποτελεσματικότητα της μεθόδου και τη σημασία του διανύσματος της πρόσφατης προσοχής Π.
- Συγκρίνουμε την αποτελεσματικότητα της μεθόδου σε σχέση με τις άλλες τεχνολογίες αιχμής.
- Εξετάζουμε το ρυθμό σύγκλισης της μεθόδου.

Στην πειραματική μας αξιολόγηση χρησιμοποιούμε τα ίδια σύνολα δεδομένων που χρησιμοποιήθηκαν και στο Κεφάλαιο 4. Επιπλέον ακολουθούμε την πειραματική διαδικασία του Κεφαλαίου 4 κάνοντας διάτμηση των συνόλων δεδομένων σε μέλλουσα και τρέχουσα κατάσταση. Όπως και στο Κεφάλαιο 4 χρησιμοποιούμε ως μετρικές αξιολόγησης το $\rho$ του Spearman και το nDCG. Οι τιμές βάσης για το λόγο $\eta$, του μεγέθους της μέλλουσας προς την τρέχουσα κατάσταση, καθώς και για το σύνολο $k$ στο οποίο εξετάζουμε το nDCG είναι επίσης ίδιες με αυτές του Κεφαλαίου 4.

[14] Αυτό ισχύει καθώς με τη χρήση των βαθμολογιών $T(p_i)$ οι οποίες δεν είναι ποτέ μηδενικές, έχουμε σύνδεση μεταξύ όλων των κόμβων.

Πίνακας 5.2: Χώρος παραμετροποίησης της μεθόδου μας

| **Παράμετρος** | **Ελάχιστο** | **Μέγιστο** | **Βήμα** |
|---|---|---|---|
| $\alpha$ | 0.0 | 0.5 | 0.1 |
| $\beta$ | 0.0 | 1.0 | 0.1 |
| $\gamma$ | 0.0 | 0.9 | 0.1 |
| $y$ | 1 | 5 | 1 |

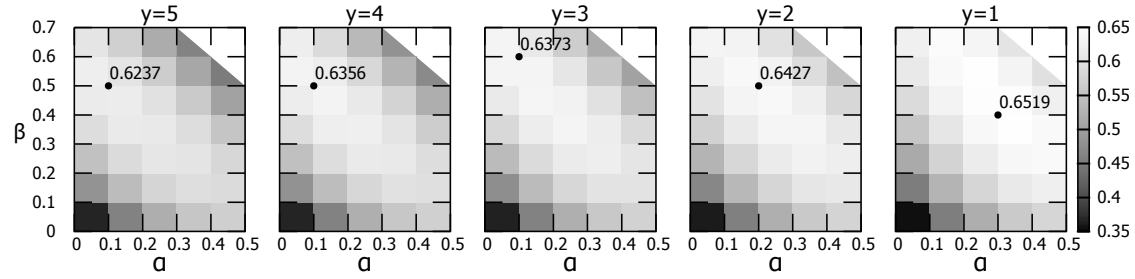



Σχήμα 5.7: Εμπειρική κατανομή της πιθανότητας να λάβει αναφορά μια δημοσίευση $n$ έτη μετά την έκδοσή της ($n \leq 10$). Η κατανομή προκύπτει από την ανάλυση των δεδομένων για τα 4 σύνολα δεδομένων που χρησιμοποιούμε.

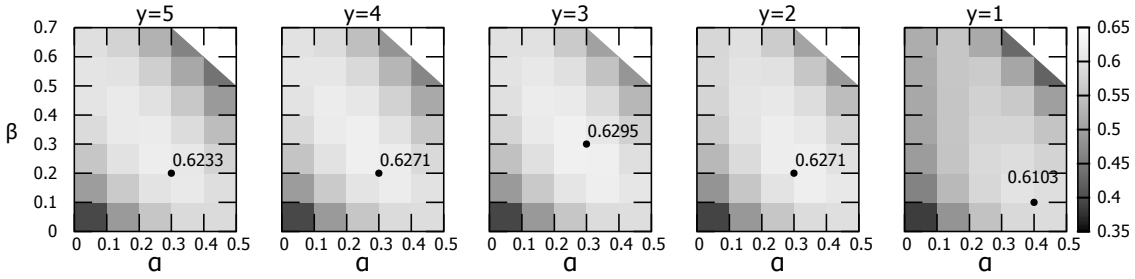
#### 5.2.4.1 Αποτελεσματικότητα Κατατάξεων της Μεθόδου Μας

Εξετάζουμε εδώ την αποτελεσματικότητα της μεθόδου μας θέτοντας το λόγο $\eta$ και το σύνολο $k$ στις τιμές βάσης τους ($\eta = 1.6, k = 50$), μεταβάλλοντας τις τιμές των παραμέτρων $\alpha, \beta, \gamma$ και $y$. Το εύρος τιμών που εξετάζουμε παρουσιάζεται στον Πίνακα 5.2. Στη συνέχεια αναζητούμε για κάθε μετρική αποτελεσματικότητας τη βέλτιστη ρύθμιση των παραμέτρων.

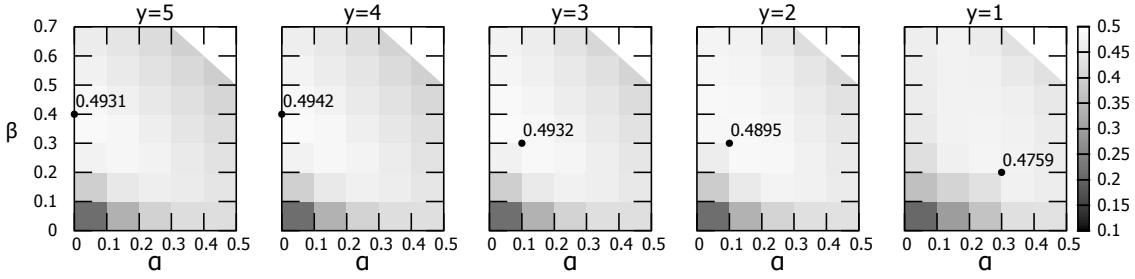
Πριν προχωρήσουμε σ' αυτό όμως, πρώτα περιγράφουμε πως θέτουμε την τιμή του εκθετικού παράγοντα $w$ της Εξίσωσης 5.6. Ακολουθούμε μια διαδικασία σαν αυτή που ακολουθείται στο [66]. Για κάθε σύνολο δεδομένων χρησιμοποιούμε μια εκθετική συνάρτηση της μορφής $e^{\widetilde{w}y}$, την οποία ταιριάζουμε στην κατανομή της τυχαίας μεταβλητής $Y$ που μοντελοποιεί την πιθανότητα μια δημοσίευση να αναφερθεί $n$ έτη μετά από την έκδοσή της. Στο Σχήμα 5.7 παρουσιάζεται αυτή η εμπειρική κατανομή για τα σύνολα δεδομένων που χρησιμοποιούμε. Ο παράγοντας $\widetilde{w}$ της συνάρτησης προσαρμογής στα δεδομένα χρησιμοποιείται ως τιμή για τον παράγοντα $w$. Ακολουθώντας αυτή τη διαδικασία υπολογίζουμε $w = -0.48$ για το σύνολο hep-th, $w = -0.12$ για το σύνολο APS και $w = -0.16$ για τα σύνολα PMC και DBLP.



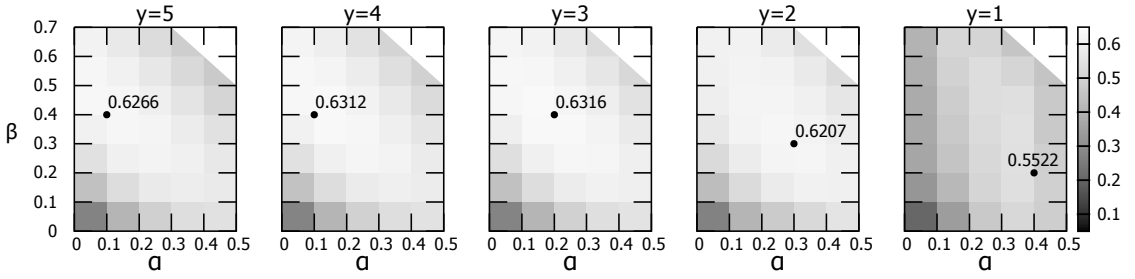

(α′) hep-th.




(β′) APS.




(γ′) PMC.




(δ′) DBLP.

Σχήμα 5.8: Διαγράμματα θερμότητας που απεικονίζουν τα αποτελέσματα της κάθε παραμετροποίησης της μεθόδου μας βάσει της συσχέτισης με το υπόβαθρο αληθείας για κάθε σύνολο δεδομένων. Η τιμή που επιτυγχάνεται για τη βέλτιστη παραμετροποίηση (Spearman's $\rho$) σημειώνεται στο κάθε σχήμα.

**Αποτελεσματικότητα με βάση τη Συσχέτιση.** Σε αυτό το πείραμα μετράμε την αποτελεσματικότητα παραγωγής κατατάξεων της μεθόδου μας, χρησιμοποιώντας ως μετρική το $\rho$ του Spearman ως προς το υπόβαθρο αληθείας P-CC. Οπτικοποιούμε την αποτελεσματικότητα κάθε παραμετροποίησης που εξετάσαμε μέσω ενός χάρτη θερμότητας (Heatmap) που απεικονίζει το χώρο $\alpha$–$\beta$ για διαφορετικές τιμές του $y$.[15] Τα αποτελέσματα για τις διάφορες τιμές παραμέτρων φαίνονται στα Σχήματα 5.8α΄-5.8δ΄

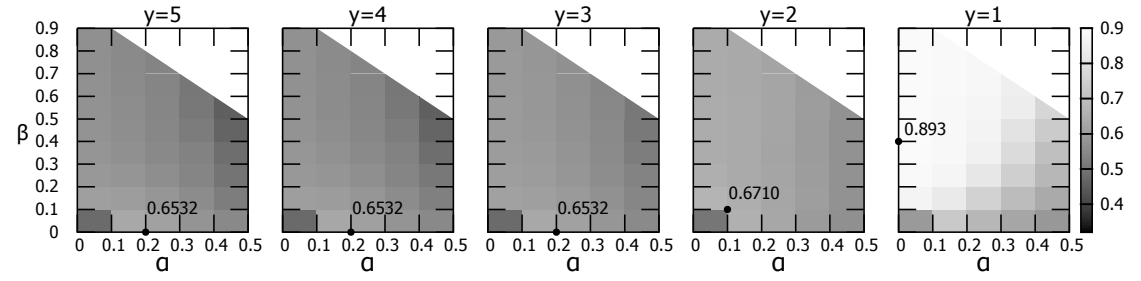
Στους χάρτες θερμότητας βλέπουμε τα αποτελέσματα για τις διάφορες τιμές των παραμέτρων $\alpha$ και $\beta$. Καθώς το $\alpha$ αυξάνει η μέθοδος μας προσομοιώνει ερευνητές που κυρίως προτιμούν να διαβάζουν δημοσιεύσεις από λίστες αναφορών άλλων δημοσιεύσεων και σπανιότερα επιλέγουν να διαβάζουν δημοσιεύσεις βάσει της ηλικίας τους, ή της πρόσφατης προσοχής τους. Καθώς $\alpha \to 1$, επομένως, η μέθοδος μας σταδιακά εκφυλίζεται προς το απλό PageRank, με μικρή πιθανότητα τυχαίων μεταβάσεων. Συνεπώς, αναμένουμε με την αύξηση του $\alpha$ να πέφτει η συσχέτιση της κατάταξης της μεθόδου με το υπόβαθρο αληθείας. Αυτό επιβεβαιώνεται από το γεγονός ότι για $\alpha = 0.5$ σε όλα τα διαγράμματα παρατηρούμε πιο σκούρα χρώματα σε σχέση με $\alpha = 0.4$ ενώ η συσχέτιση πέφτει ακόμα περισσότερο για μεγαλύτερες τιμές του. Επιπλέον, οι χάρτες θερμότητας επιβεβαιώνουν τη σημασία του διανύσματος της πρόσφατης προσοχής, καθώς για $\beta = 0$ (NO-ATT) παρατηρούμε σημαντικά χαμηλότερες συσχετίσεις (σκούρες αποχρώσεις στην κάτω αριστερά γωνία των διαγραμμάτων). Παρομοίως, παρατηρήσαμε κάπως χαμηλότερες συσχετίσεις όταν $\beta = 1$ (ATT-ONLY).

Συνολικά παρατηρούμε ότι οι κατατάξεις που παράγει η μέθοδος μας, στην καλύτερη παραμετροποίησή της συσχετίζονται τουλάχιστον μέτρια με την κατάταξη του υπόβαθρου αληθείας για όλα τα σύνολα δεδομένων ($\rho > 0.49$). Επιπλέον παρατηρούμε ότι η βέλτιστη τιμή για τον αριθμό ετών που χρησιμοποιούμε στον υπολογισμό του διανύσματος προσοχής είναι $y = \{1, 3, 3, 4\}$ για τα σύνολα {hep-th, APS, PMC, DBLP} αντίστοιχα. Ιδιαίτερο ενδιαφέρον έχει το γεγονός ότι τα τρία τελευταία σύνολα δεδομένων ακολουθούν παρόμοια μοτίβα όσον αφορά την κατανομή των αναφορών (Σχήμα 5.7), με τις περισσότερες αναφορές να λαμβάνονται $2-3$ χρόνια μετά την έκδοση των δημοσιεύσεων, ενώ στο σύνολο hep-th αυτό συμβαίνει νωρίτερα. Διαισθητικά, επομένως, είναι λογικό να χρησιμοποιούμε μικρότερες τιμές για το $y$ για να υπολογίσουμε το διάνυσμα προσοχής στο hep-th. Οι τάσεις στην έρευνα του συγκεκριμένου συνόλου φαίνεται να αλλάζουν γρηγορότερα και άρα ένα μεγαλύτερο χρονικό παράθυρο θα έδινε μια εικόνα για παρελθοντικές και όχι τρέχουσες τάσεις. Στα σύνολα APS, PMC και DBLP από την άλλη, οι δημοσιεύσεις αποκτούν αναφορές με πιο αργό ρυθμό, επομένως μεγαλύτερες τιμές για το $y$ είναι πιο πιθανό να αντανακλούν τις τρέχουσες τάσεις στην έρευνα.

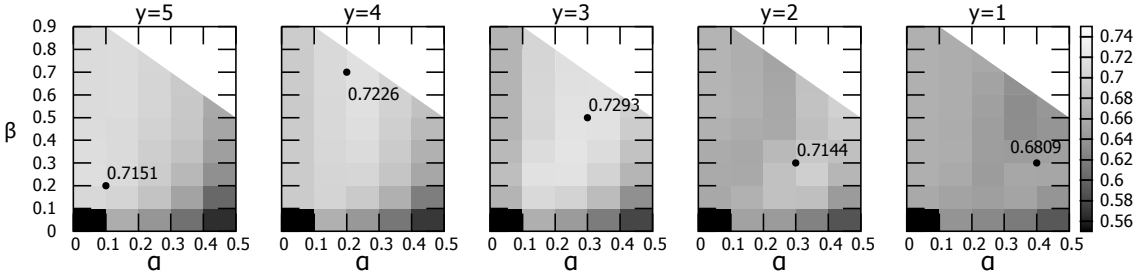
Από τους χάρτες θερμότητας μπορούμε επίσης να καθορίσουμε τις παραμετροποιήσεις που επιτυγχάνουν τη μέγιστη συσχέτιση με το υπόβαθρο αληθείας. Συγκεκριμένα, οι τιμές που βρίσκουμε για τις παραμέτρους $\{\alpha, \beta, \gamma, y\}$ είναι $\{0.3, 0.4, 0.3, 1\}$ για το hep-th ($\rho = 0.6519$), $\{0.3, 0.3, 0.4, 3\}$ για το APS ($\rho = 0.6295$), $\{0.0, 0.4, 0.6, 4\}$ για το PMC ($\rho = 0.494$), και $\{0.2, 0.4, 0.4, 3\}$ για το DBLP ($\rho = 0.6316$).

Για να αναδείξουμε τη σημασία του διανύσματος προσοχής που εισαγάγαμε, συγκρίνουμε τα αποτελέσματα αυτά με τις μέγιστες τιμές που επιτυγχάνονται όταν $\beta = 0$ (NO-ATT). Αυτές είναι 0.56, 0.581, 0.411, και 0.529 για τα hep-th, APS, PMC, και DBLP, αντίστοιχα. Επιπλέον για $\beta = 1$, οι τιμές αυτές είναι 0.615, 0.537, 0.45, 0.571.

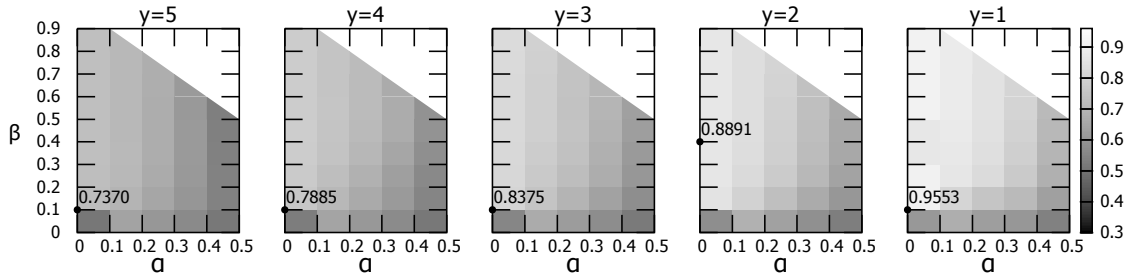
[15] Η τιμή του $\gamma$ υπονοείται καθώς $\alpha + \beta + \gamma = 1$.



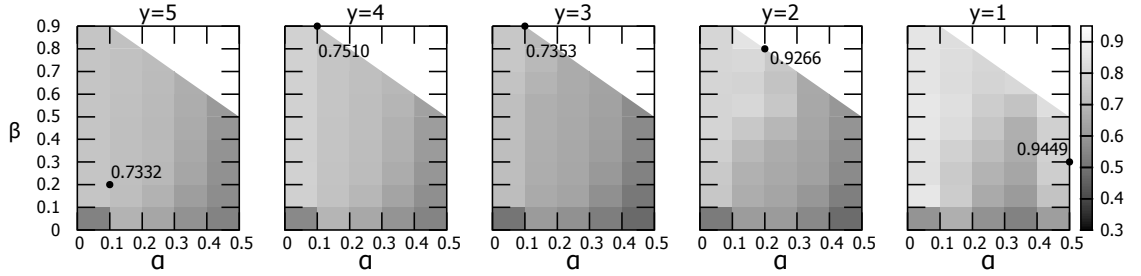

(α′) hep-th.




(β′) APS.




(γ′) PMC.




(δ′) DBLP.

Σχήμα 5.9: Διαγράμματα θερμότητας που απεικονίζουν τα αποτελέσματα της κάθε παραμετροποίησης της μεθόδου μας βάσει του nDCG@50 με το υπόβαθρο αληθείας για κάθε σύνολο δεδομένων. Η τιμή που επιτυγχάνεται για τη βέλτιστη παραμετροποίηση σημειώνεται στο κάθε σχήμα.

**Αποτελεσματικότητα ως προς το nDCG@50**. Επαναλαμβάνουμε την ανάλυση γύρω από την αποτελεσματικότητα της μεθόδου μας, μετρώντας την αυτή τη φορά μέσω της μετρικής nDCG ($k = 50$). Παρουσιάζουμε τους χάρτες θερμότητας στα Σχήματα 5.9α΄-5.9δ΄.

Μια πρώτη ενδιαφέρουσα παρατήρηση είναι ότι όσον αφορά την αναγνώριση των πιο δημοφιλών δημοσιεύσεων φαίνεται να ευνοεί η χρήση μικρότερων χρονικών διαστημάτων στα οποία υπολογίζονται οι τιμές προσοχής. Παρατηρούμε ότι καθώς αυξάνει το $y$ η τιμές του nDCG φθίνουν γρήγορα, (παρατηρούμε πιο σκούρες αποχρώσεις για $y > 1$ στο σύνολο των διαγραμμάτων). Είναι αναμενόμενο όσο αυξάνει το $y$ οι τιμές του nDCG να φθίνουν ακόμα περισσότερο, επειδή αυξάνοντας το χρονικό διάστημα $y$ επαναεισάγουμε την εγγενή μεροληψία των δικτύων απέναντι σε νεότερες δημοσιεύσεις. Έτσι οι δημοσιεύσεις με τις μεγαλύτερες τιμές Π δεν σχετίζονται πλέον με τις τρέχουσες τάσεις στην έρευνα. Το ίδιο ισχύει για αυξημένες τιμές του $\alpha$ όταν $y > 1$. Καθώς αυξάνει το $\alpha$ η λογική του απλού PageRank κυριαρχεί, ευνοώντας έτσι περισσότερο τις πιο παλιές δημοσιεύσεις. Αυτές σε μεγάλο βαθμό δεν θα είναι πλέον το σημείο στο οποίο επικεντρώνεται η τρέχουσα έρευνα. Η παρατήρηση αυτή προκύπτει από τις πιο σκούρες αποχρώσεις σε όλα τα διαγράμματα όταν το $\alpha$ προσεγγίζει την τιμή 0.5.

Τέλος, καθορίζουμε την παραμετροποίηση που επιτυγχάνει το καλύτερο nDCG ανά σύνολο δεδομένων. Συγκεκριμένα υπολογίζουμε ότι οι καλύτερες τιμές για τις παραμέτρους {$\alpha$, $\beta$, $\gamma$, $y$} σε αυτό το σενάριο είναι {0.0, 0.4, 0.6, 1} για το hep-th ($nDCG = 0.8930$), {0.3, 0.5, 0.2, 3} για το σύνολο APS ($nDCG = 0.7293$), {0.0, 0.1, 0.9, 1} για το σύνολο PMC ($nDCG = 0.9553$) και {0.5, 0.3, 0.2, 1} για το σύνολο DBLP ($nDCG = 0.9449$). Όπως και στο προηγούμενο πείραμα, βρίσκουμε ότι το διάνυσμα προσοχής έχει μη αμελητέο ρόλο στην επίτευξη των βέλτιστων τιμών nDCG σε όλα τα σύνολα δεδομένων ($\beta > 0$). Ενδεικτικά, οι μέγιστες τιμές nDCG για την περίπτωση όπου $\beta = 0$ είναι 0.669, 0.635, 0.6, και 0.663 για τα hep-th, APS, PMC, και DBLP, αντίστοιχα. Για $\beta = 1$ οι αντίστοιχες τιμές είναι 0.89, 0.692, 0.916, 0.916.

#### 5.2.4.2 Συγκριτική Αξιολόγηση της Μεθόδου Μας

**Συγκριτική Αξιολόγηση ως προς τη συσχέτιση.** Στις επόμενες παραγράφους συγκρίνουμε την αποτελεσματικότητα της μεθόδου μας με άλλες τεχνολογίες αιχμής. Συγκεκριμένα, επιλέγουμε τις μεθόδους που είχαν τη βέλτιστη αποτελεσματικότητα παραγωγής κατατάξεων με βάση την δημοφιλία, όπως προέκυψε από την πειραματική αξιολόγηση στο Κεφάλαιο 4 (CR, FR, RAM, ECM), καθώς και την πιο πρόσφατη από αυτές που αξιολογήσαμε (WSDM). Δεδομένου ότι για τη μέθοδο μας πραγματοποιήσαμε διεξοδικά πειράματα παραμετροποίησης και για να είναι ορθή η σύγκριση της αποτελεσματικότητάς της σε σχέση με αυτή των ανταγωνιστριών μεθόδων, πραγματοποιήσαμε και για τις τελευταίες μια εξαντλητική εξέταση των παραμέτρων τους. Παρουσιάζουμε στον Πίνακα 5.3 το εύρος των τιμών και το βήμα για τις παραμέτρους τους που εξετάσαμε. Η μέθοδος WSDM δεδομένου ότι χρησιμοποιεί δεδομένα για τα περιοδικά στα οποία εκδόθηκαν οι δημοσιεύσεις, εφαρμόστηκε μόνο στα σύνολα δεδομένων PMC και DBLP, για τα οποία η πληροφορία αυτή ήταν διαθέσιμη. Επιπλέον, όπως και στο Κεφάλαιο 4, χρησιμοποιήσαμε για όλες τις επαναληπτικές μεθόδους σφάλμα σύγκλισης $\epsilon = 10^{-12}$. Επιπλέον για να αναδείξουμε τη σημασία του διανύσματος πρόσφατης προσοχής που εισαγάγαμε, συμπεριλαμβάνουμε στην αξιολόγηση και τις περιπτώσεις όπου στη μέθοδο μας έχει τεθεί $\beta = 0$ (NO-ATT) και

Πίνακας 5.3: Χώρος παραμετροποίησης των ανταγωνιστών.

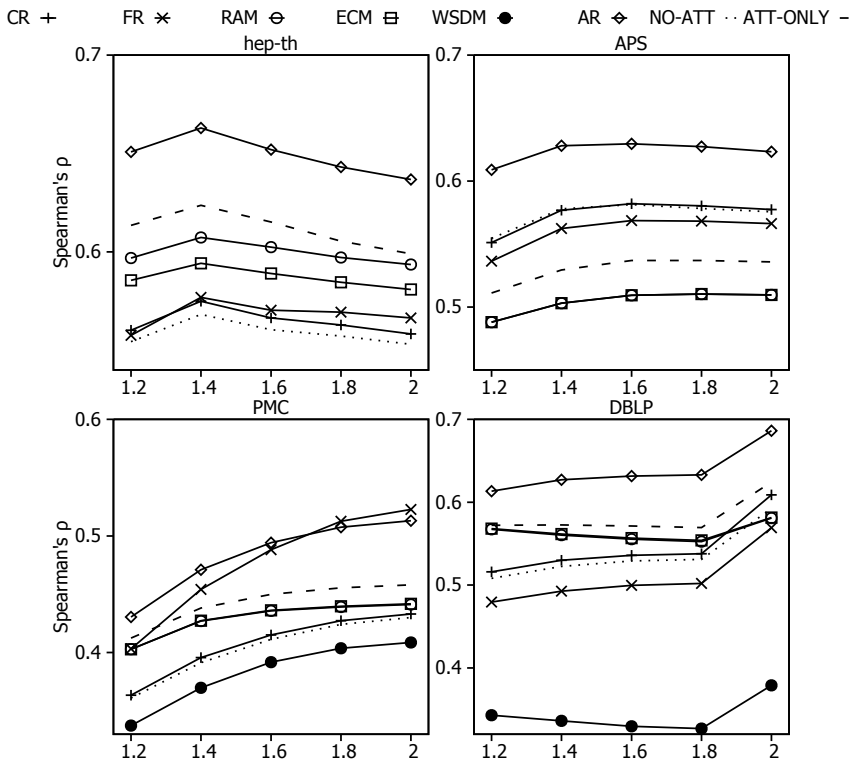
| **Μέθοδος** | **Παράμετρος** | **Ελάχιστο** | **Μέγιστο** | **Βήμα** |
|---|---|---|---|---|
| **CR** | $\alpha$ | 0.1 | 0.7 | 0.2 |
| | $\tau_{dir}$ | 2 | 10 | 2 |
| **FR** | $\alpha$ | 0.1 | 0.5 | 0.1 |
| | $\beta$ | 0.0 | 0.9 | 0.1 |
| | $\gamma$ | 0.0 | 0.9 | 0.1 |
| | $\rho$ | -0.82 | -0.42 | 0.2 |
| **RAM** | $\gamma$ | 0.1 | 0.9 | 0.1 |
| **ECM** | $\alpha$ | 0.1 | 0.5 | 0.1 |
| | $\gamma$ | 0.1 | 0.5 | 0.1 |
| **WSDM** | $\alpha$ | 1.1 | 2.3 | 0.3 |
| | $\beta$ | 1 | 5 | 1 |
| | $i$ | 4 | 5 | 1 |

$\beta = 1$ (ATT-ONLY).

Η πειραματική διαδικασία που ακολουθούμε ακολουθεί τα πρότυπα αυτής του Κεφαλαίου 4, όπου μεταβάλλουμε το λόγο $\eta$. Για κάθε σύνολο δεδομένων και για κάθε τιμή του $\eta$ χρησιμοποιούμε εκείνη την παραμετροποίηση κάθε μεθόδου που έδωσε τα καλύτερα αποτελέσματα, μετρώντας τη συσχέτιση της κατάταξης που παράγει με το υπόβαθρο αληθείας P-CC. Τα αποτελέσματα παρουσιάζονται στο Σχήμα 5.10. Η μέθοδος που προτείνουμε σημειώνεται στο εξής στα διαγράμματα των πειραμάτων στην παρούσα ενότητα και σε αυτές που ακολουθούν ως $AR$.

Παρατηρούμε ότι η μέθοδος μας επιτυγχάνει καλύτερη συσχέτιση σε σύγκριση με όλες τις ανταγωνίστριες μεθόδους και για όλες τις μετρήσεις στα σύνολα δεδομένων hep-th, APS και DBLP. Συγκεκριμένα η μέθοδός μας επιτυγχάνει βελτιωμένη συσχέτιση εώς και κατά 0.055 στο σύνολο δεδομένων hep-th, κατά 0.057 στο σύνολο APS και μέχρι και 0.077 βελτιωμένη συσχέτιση στο σύνολο DBLP σε σύγκριση με την καλύτερη ανταγωνίστρια μέθοδο. Επιπλέον, στις περισσότερες περιπτώσεις η μέθοδος μας υπερισχύει των ανταγωνιστών στο σύνολο PMC έως και κατά 0.027 στην συσχέτιση που επιτυγχάνεται, ενώ οριακά δεν υπερισχύει της μεθόδου FR σε δύο σημεία μετρήσεων (η διαφορά που παρατηρείται σε αυτές τις περιπτώσεις στη συσχέτιση είναι ωστόσο μόλις της τάξης του 0.01). Αξίζει να σημειωθεί εδώ ότι η μέθοδος FR που υπερισχύει έναντι της μεθόδου μας σε αυτό το σύνολο δεδομένων, επιτυγχάνει καλά αποτελέσματα μόνο σε αυτό το σύνολο. Αντίθετα, στα υπόλοιπα σύνολα η μέθοδος FR δεν υπερισχύει έναντι των άλλων. Από την άλλη, η μέθοδος μας σταθερά υπερισχύει έναντι των άλλων μεθόδων σε όλα τα σύνολα δεδομένων, επιτυγχάνοντας μάλιστα αρκετά βελτιωμένη συσχέτιση.

Η αποτελεσματικότητα της μεθόδου μας οφείλεται στο γεγονός ότι, σε σχέση με τις ανταγωνίστριες που χρησιμοποιούν χρονικούς παράγοντες, δεν προωθεί απλώς δημοσιεύσεις που πήραν αναφορές, ή δημοσιεύτηκαν πρόσφατα. Αντ' αυτού, λόγω του μηχανισμού της πρόσφατης προσοχής, προωθεί εντονότερα τις πρόσφατες δημοσιεύσεις που έλαβαν και πολλές αναφορές, σε σχέση με τις πρόσφατες δημοσιεύσεις που δεν έχουν λάβει πολλές αναφορές. Όπως αναφέραμε και στην Ενότητα 5.2.1, δημοσιεύσεις που υπήρξαν δημοφιλείς πρόσφατα πράγματι παραμένουν δημοφιλείς σε μεγάλο ποσοστό. Επιπλέον η μέθοδός μας προωθεί και παλιές δημοσιεύσεις που εξακολουθούν να είναι στο επίκεντρο της προσοχής. Η σημασία του μηχανισμού προσοχής αναδεικνύεται




Σχήμα 5.10: Αποτελεσματικότητα όλων των μεθόδων όσον αφορά τη συνολική συσχέτιση ($\rho$ του Spearman), καθώς μεταβάλλουμε το λόγο $\eta$ (άξονας X).

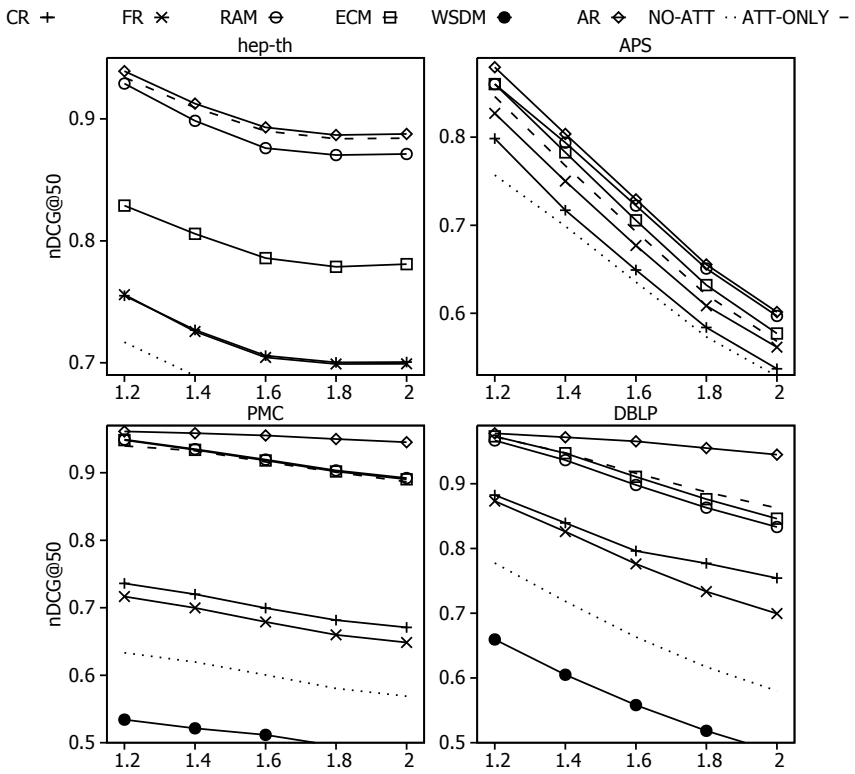
από το γεγονός ότι σε δύο σύνολα δεδομένων η περίπτωση ATT-ONLY είναι πιο αποτελεσματική από τις υπόλοιπες τεχνολογίες αιχμής. Από την άλλη, αγνοώντας πλήρως το μηχανισμό της προσοχής (NO-ATT) πετυχαίνουμε αισθητά χαμηλότερη αποτελεσματικότητα. Σε κάθε περίπτωση η αποτελεσματικότητα αυξάνει όταν ο μηχανισμός της πρόσφατης προσοχής συνδυάζεται με τους άλλους δύο μηχανισμούς.

**Συγκριτική αξιολόγηση ως προς το nDCG.** Σε αυτή τη παράγραφο μετράμε την τιμή nDCG που επιτυγχάνει κάθε μέθοδος με βάση το υπόβαθρο αληθείας P-CC. Πραγματοποιούμε δύο πειράματα: στο πρώτο πείραμα θέτουμε $k = 50$ για τον υπολογισμό του nDCG, μεταβάλλοντας το λόγο $\eta$. Στο δεύτερο πείραμα χρησιμοποιούμε την τιμή βάσης $\eta = 1.6$ και μετράμε το nDCG μεταβάλλοντας την τιμή του $k$.

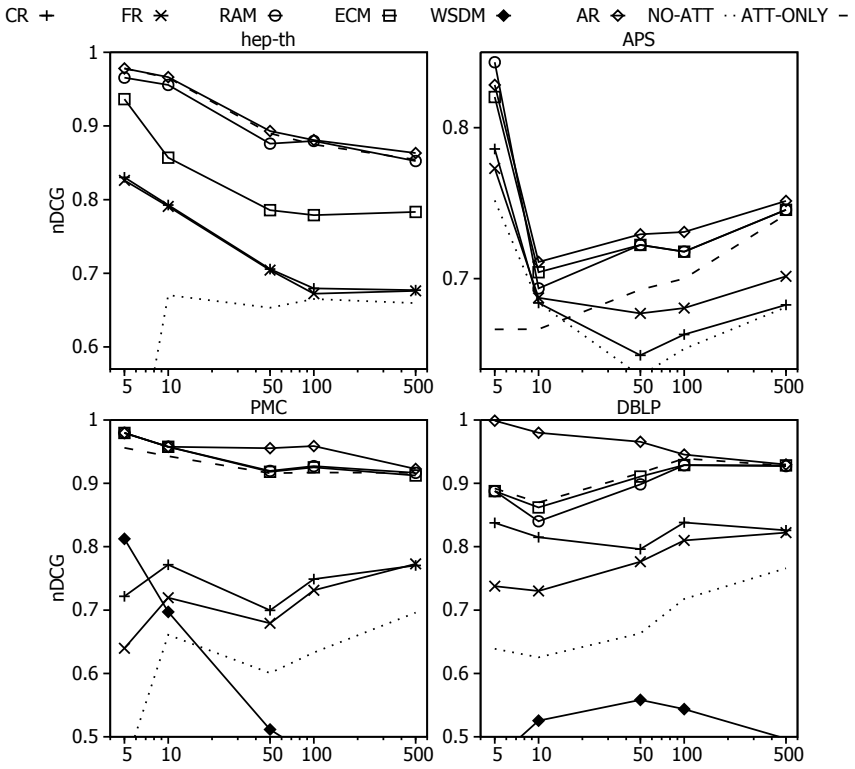
Στο Σχήμα 5.11 παρουσιάζουμε τα αποτελέσματα του πρώτου πειράματος. Για κάθε σημείο στο οποίο παίρνουμε μετρήσεις παρουσιάζουμε την τιμή που προκύπτει για το εκάστοτε βέλτιστο σύνολο παραμέτρων κάθε μεθόδου.

Γενικά, καθώς κοιτάμε πιο μακριά στο μέλλον (αυξάνει το $\eta$) η ακρίβεια των παραγόμενων κατατάξεων μειώνεται για όλες τις μεθόδους. Το φαινόμενο αυτό είναι πιο έντονο στο σύνολο δεδομένων APS και λιγότερο στο hep-th. Σε όλες τις περιπτώσεις η μέθοδος μας είναι αποτελεσματικότερη από τις ανταγωνίστριες μεθόδους, ενώ η βελτίωση που επιτυγχάνεται αυξάνει με την αύξηση του $\eta$ σε δύο από τα σύνολα δεδομένων. Συγκεκριμένα, η μέθοδός μας πετυχαίνει καλύτερες τιμές nDCG ως και κατά 0.017 στο σύνολο hep-th, 0.018 στο APS, 0.053 στο PMC και 0.098 στο DBLP σε σύγκριση με την πιο αποτελεσματική ανταγωνίστρια μέθοδο. Σημειώνεται ότι οι πιο αποτελεσματικές ανταγωνίστριες μέθοδοι είναι οι RAM και ECM.

Στο Σχήμα 5.12 παρουσιάζουμε τα αποτελέσματα του δεύτερου πειράματος, όπου μεταβάλλουμε την τιμή του $k$ διατηρώντας σταθερό το λόγο $\eta$ στην τιμή $\eta = 1.6$. Σε κάθε σημείο μετρήσεων και πάλι επιλέγουμε για κάθε μέθοδο την παραμετροποίηση




Σχήμα 5.11: Αποτελεσματικότητα όλων των μεθόδων με βάση το nDCG@50. Στον άξονα X παρουσιάζονται οι τιμές του λόγου $\eta$.




Σχήμα 5.12: Αποτελεσματικότητα όλων των μεθόδων ως προς το nDCG@$k$ με το λόγο $\eta$ στην τιμή βάσης του ($\eta = 1.6$). Ο άξονας X αντιστοιχεί στις τιμές του $k$.

που δίνει το καλύτερο αποτέλεσμα. Γενικά παρατηρούμε ότι η μέθοδος μας είναι, στην χειρότερη περίπτωση, εξίσου αποτελεσματική με την καλύτερη ανταγωνίστρια, ενώ στις

περισσότερες περιπτώσεις είναι αποτελεσματικότερη από όλες τις ανταγωνίστριες. Η μόνη εξαίρεση παρατηρείται για το σύνολο APS και για $k = 5$ (όπου ωστόσο η διαφορά με την καλύτερη ανταγωνίστρια μέθοδο είναι 0.015). Συγκεκριμένα, η μέθοδος μας πετυχαίνει καλύτερες τιμές nDCG ως και κατά 0.017 στο σύνολο δεδομένων hep-th, 0.013 στο APS (εξαίρεση για $k = 5$), 0.035 στο PMC και 0.111 στο DBLP. Επιπλέον, για μικρές τιμές του $k$ ($k = \{5, 10\}$) η μέθοδός μας πετυχαίνει τιμές κοντά στο 1 σε τρία από τα τέσσερα σύνολα δεδομένων (στα hep-th, PMC και DBLP). Οι καλύτερες ανταγωνίστριες μέθοδοι είναι και πάλι οι RAM και ECM, ανάλογα με το σύνολο δεδομένων.

Στις ειδικές περιπτώσεις της μεθόδου μας (NO-ATT, ATT-ONLY), παρατηρούμε στα Σχήματα 5.11 και 5.12 ότι αν δεν χρησιμοποιήσουμε το διάνυσμα της πρόσφατης προσοχής έχουμε σημαντική πτώση στις τιμές nDCG που επιτυγχάνονται. Από την άλλη, χρησιμοποιώντας μόνο το διάνυσμα προσοχής, σε πολλές περιπτώσεις πετυχαίνουμε καλύτερα αποτελέσματα από τις ανταγωνίστριες μεθόδους, με εξαίρεση την περίπτωση του συνόλου APS. Όπως συμβαίνει και στην περίπτωση του Σχήματος 5.10, ο συνδυασμός των τριών μηχανισμών της μεθόδου μας δίνει σημαντική βελτίωση στην ακρίβεια των παραγόμενων κατατάξεων.

**Σύγκλιση της Μεθόδου Μας.** Η μέθοδος μας περιλαμβάνει μια επαναληπτική διαδικασία, παρόμοια με αυτή του PageRank, για να υπολογίσει τις βαθμολογίες των δημοσιεύσεων. Συγκεκριμένα μπορούμε να θεωρήσουμε την μέθοδο μας ως παραλλαγή του PageRank όπου το διάνυσμα τυχαίας μετάβασης αντικαθίσταται από δύο διανύσματα: το διάνυσμα προσοχής και το διάνυσμα που εξαρτάται από τις ηλικίες των δημοσιεύσεων. Συνεπώς, στη μέθοδο μας η πιθανότητα τυχαίας μετάβασης $1 - \alpha$ του PageRank μοιράζεται στις πιθανότητες $\beta, \gamma$. Η σύγκλιση της μεθόδου επομένως εξαρτάται από τους ίδιους παράγοντες που επηρεάζουν τη σύγκλιση του PageRank. Οι ιδιότητες αυτές (καθώς και γενικότερα οι ιδιότητες του PageRank) παρουσιάζονται αναλυτικά στο [47]. Το πιο σημαντικό στοιχείο είναι ότι καθώς $\alpha \to 1$ η ταχύτητα σύγκλισης μειώνεται και απαιτούνται περισσότερες επαναλήψεις. Ωστόσο, όπως είδαμε στις προηγούμενες παραγράφους, οι μεγάλες τιμές για το $\alpha$ δεν ευνοούν την αποτελεσματική παραγωγή κατατάξεων βάσει της δημοφιλίας, ενώ η βέλτιστη αποτελεσματικότητα της μεθόδου μας επιτυγχάνεται όταν $\alpha \leq 0.5$. Επιπλέον, καθώς $\alpha \to 0$, η μέθοδος τείνει να εξαρτάται όλο και περισσότερο μόνο από το άθροισμα των διανυσμάτων της προσοχής και ηλικίας των δημοσιεύσεων. Επομένως ο αριθμός των επαναλήψεων που απαιτείται μειώνεται, με την οριακή περίπτωση $\alpha = 0$ να απαιτεί μόνο μια επανάληψη.

Συνολικά αναμένουμε η μέθοδος μας να συγκλίνει πιο γρήγορα από το PageRank και τις παραλλαγές του (το PageRank συνήθως στα δίκτυα αναφορών εφαρμόζεται με $\alpha = 0.5$ [15, 54]). Στα πειράματά μας, η μέθοδος μας συγκλίνει σε λιγότερες από 30 επαναλήψεις στα σύνολα hep-th, APS, DBLP και σε λιγότερες από 20 επαναλήψεις στο PMC για $\alpha = 0.5$ και σφάλμα σύγκλισης $\epsilon \leq 10^{-12}$, ενώ ο αριθμός επαναλήψεων που απαιτείται είναι μικρότερος για μικρότερες τιμές του $\alpha$. Αντίθετα, οι απαιτούμενες επαναλήψεις για τη μέθοδο CR είναι 51, 46, 26, και 47, στα σύνολα hep-th, APS, PMC, και DBLP, αντίστοιχα, για $\alpha = 0.5$. Αντίστοιχα για τη μέθοδο FR (η οποία στη πράξη δεν συγκλίνει σε κάθε περίπτωση) απαιτούνταν 35, 30, 26, και 23 επαναλήψεις στα σύνολα hep-th, APS, PMC, και DBLP, αντίστοιχα, για $\alpha = 0.5$.

## 5.3 Συμπεράσματα

Σε αυτή την ενότητα παρουσιάσαμε πραγματικά συστήματα που κάνουν χρήση αποτελεσματικών μεθόδων κατάταξης, που έχουν βασιστεί στην πειραματική διαδικασία που αναπτύχθηκε στο Κεφάλαιο 4, ώστε να εμφανίζουν αποτελέσματα αναζήτησης με βάση τη δημοφιλία ή την επιρροή. Το σύστημα BIP! είναι μια ακαδημαϊκή μηχανή αναζήτησης γενικού σκοπού που παρέχει πολλές χρήσιμες πληροφορίες αναλυτικής φύσης. Η καινοτομία του είναι ότι δίνει στο χρήστη τη δυνατότητα να επιλέγει το κριτήριο βάσει του οποίου γίνεται η κατάταξη των αποτελεσμάτων (δημοφιλία/επιρροή), ενώ συνδυάζει τις κατατάξεις αυτές και με τη σχετικότητα των κειμένων με την εκάστοτε αναζήτηση. Τέλος, με το ελεύθερα διαθέσιμο API του δίνει κίνητρο για την ανάπτυξη επιπλέον εφαρμογών. Από την άλλη το mirPub v2 επεκτείνει τη μηχανή αναζήτησης mirPub, η οποία εστιάζει στην αναζήτηση δημοσιεύσεων του κλάδου της μοριακής βιολογίας που σχετίζονται με miRNAs. Το σύστημα διευκολύνει τους βιολόγους ερευνητές να βρίσκουν δημοσιεύσεις που τους ενδιαφέρουν, κάνοντας χρήση μηχανισμών επέκτασης λέξεων-κλειδιών και μεθόδων κατάταξης που έχουν παραμετροποιηθεί ώστε να κατατάσσουν τις δημοσιεύσεις με βάση την επιρροή τους.

Ακόμα, μελετήσαμε στο τρέχον κεφάλαιο και το πρόβλημα της κατάταξης δημοσιεύσεων με βάση τη δημοφιλία. Καρπός αυτής της μελέτης ήταν ο σχεδιασμός και η υλοποίηση μιας νέας μεθόδου που προτείναμε. Κύριο συμπέρασμα της εργασίας μας είναι η σημασία του διανύσματος προσοχής, το οποίο εντοπίζει δημοφιλείς δημοσιεύσεις, οι οποίες είναι πιθανό να εξακολουθήσουν να λαμβάνουν πολλές αναφορές στο κοντινό μέλλον. Μελετήσαμε την αποτελεσματικότητα της μεθόδου σε τέσσερα πραγματικά σύνολα δεδομένων και δείξαμε ότι είναι αποτελεσματικότερη από τις τρέχουσες τεχνολογίες αιχμής στην παραγωγή κατατάξεων με βάση τη δημοφιλία, τόσο μετρώντας του $\rho$ του Spearman όσο και το μέτρου nDCG σε σχέση με το υπόβαθρο αληθείας. Επιπλέον αναδείξαμε την σημασία του διανύσματος προσοχής που εισαγάγαμε συγκρίνοντας την αποτελεσματικότητα της μεθόδου μας όταν το διάνυσμα παραλείπεται, και όταν εφαρμόζεται σκέτο, παρατηρώντας ότι σε κάθε περίπτωση ο συνδυασμός του διανύσματος με τους άλλους μηχανισμούς της μεθόδου δίνει τη βέλτιστη αποτελεσματικότητα.

# Κεφάλαιο 6

# Μελέτη Συσχέτισης Απήχησης-Αναγνωσιμότητας Δημοσιεύσεων

Σε αυτό το κεφάλαιο εξετάζουμε τη συσχέτιση της απήχησης των επιστημονικών δημοσιεύσεων με την αναγνωσιμότητα των περιλήψεων τους. Η σαφής παρουσίαση επιστημονικών θεμάτων αποτελεί θεμελιώδες κομμάτι της επιστημονικής διαδικασίας, καθώς βοηθάει στην κατανόηση των επιστημονικών ευρημάτων και θέτει τις βάσεις για μετέπειτα επιστημονική έρευνα. Επιπλέον, οι καλά γραμμένες επιστημονικές δημοσιεύσεις βοηθούν στη καλύτερη κατανόηση και διάδοση των επιστημονικών ευρημάτων από δημοσιογράφους, εκπαιδευτικούς και γενικότερα στο να φτάσει η επιστήμη στο ευρύ κοινό. Σημασία στη διαδικασία αυτή έχει η μετάδοση πληροφορίας χωρίς ανακρίβειες και παρανοήσεις.

Με δεδομένη τη σημασία της αναγνωσιμότητας, αξίζει να διερευνηθεί εάν αυτή συσχετίζεται και με την απήχηση που έχουν οι επιστημονικές δημοσιεύσεις στους ακαδημαϊκούς κύκλους. Εφόσον έχουμε αναγνωρίσει σε προηγούμενα κεφάλαια ότι η απήχηση μπορεί να έχει περισσότερες από μια διαστάσεις, αξίζει να διερευνηθεί αφενός (α) ποια άλλα ποιοτικά χαρακτηριστικά μπορούν να περιγράψουν την αξία μιας δημοσίευσης και (β) αν/πώς συσχετίζονται αυτά τα χαρακτηριστικά μεταξύ τους. Στη συνέχεια εξετάζουμε τη συσχέτιση της αναγνωσιμότητας όπως μετριέται με παραδοσιακές μετρικές, αλλά και με βάση κρίσεις ειδικών, με την απήχηση (δημοφιλία και επιρροή) των επιστημονικών δημοσιεύσεων, όπως μετριέται μέσω των βαθμολογιών μεθόδων κατάταξης που εξετάσαμε σε προηγούμενα κεφάλαια.

## 6.1 Σχετικές Εργασίες

Διάφορες εργασίες στο παρελθόν εξέτασαν την αναγνωσιμότητα επιστημονικών κειμένων και τη συσχέτισή της με την απήχηση. Οι περισσότερες εργασίες μέχρι τώρα περιορίζονται σε ένα επιστημονικό πεδίο. Για παράδειγμα στο [33] οι συγγραφείς διερευνούν τη συσχέτιση της μετρικής αναγνωσιμότητας FRE με τον αριθμό αναφορών των δημοσιεύσεων στο χώρο της ψυχολογίας. Στην εργασία αυτή οι συγγραφείς βρίσκουν να μην υπάρχει συσχέτιση μεταξύ αναγνωσιμότητας και αριθμού αναφορών στο σύνολο των δημοσιεύσεων. Ωστόσο, εξετάζοντας ένα υποσύνολο επιλεγμένων υψηλού κύρους δημοσιεύσεων του συγκεκριμένου επιστημονικού πεδίου, τα αποτελέσματα δείχνουν κάποια συσχέτιση μεταξύ αναγνωσιμότητας και απήχησης. Στο [49] εξετάζονται οι

μετρικές αναγνωσιμότητας FRE και SMOG σε ένα σύνολο δημοσιεύσεων από το αντικείμενο της επιστήμης πληροφορίας, οι οποίες εκδόθηκαν στη διάρκεια μιας δεκαετίας. Η έρευνα αυτή έδειξε να μην υπάρχει συσχέτιση μεταξύ αναγνωσιμότητας και του αριθμού αναφορών που λαμβάνουν οι δημοσιεύσεις. Στο [88] εξετάζεται η αναγνωσιμότητα δημοσιεύσεων με αντικείμενο τη νευροαπεικόνιση. Στη μελέτη αυτή επιλέχθηκαν οι 100 δημοσιεύσεις του χώρου με τις περισσότερες αναφορές και εξετάστηκαν με χρήση μιας σειράς μετρικών αναγνωσιμότητας. Αποτέλεσμα και αυτής της εργασίας ήταν να μη βρεθεί σημαντική συσχέτιση μεταξύ του αριθμού αναφορών και της αναγνωσιμότητας. Τέλος, στο [27] μελετάται το ζήτημα σε ένα διεπιστημονικό σύνολο $260,000$ περιλήψεων δημοσιεύσεων με χρήση της μετρικής SMOG. Τα αποτελέσματα αυτής της ανάλυσης επίσης έδειξαν να μην υπάρχει συσχέτιση μεταξύ της αναγνωσιμότητας και του αριθμού αναφορών.

Η εργασία μας σε αυτό το κεφάλαιο επεκτείνει τις προηγούμενες εργασίες, εξετάζοντας επιπλέον μετρικές (εξετάζουμε τέσσερις μετρικές αναγνωσιμότητας), σε ένα μεγαλύτερο σύνολο διεπιστημονικών δημοσιεύσεων. Επιπλέον, εξετάζουμε την αναγνωσιμότητα χρησιμοποιώντας τόσο παραδοσιακές μετρικές αναγνωσιμότητας, όσο και κρίσεις ειδικών όσο αφορά την αναγνωσιμότητα περιλήψεων. Τέλος, εξετάζουμε τη συσχέτιση της αναγνωσιμότητας με την απήχηση, χρησιμοποιώντας μεθόδους κατάταξης που εξετάσαμε στα κεφάλαια που προηγήθηκαν και που αντιστοιχούν σε μετρικές απήχησης (δηλαδή της δημοφιλίας και της επιρροής).

## 6.2 Μέθοδοι και Σύνολα Δεδομένων

### 6.2.1 Σύνολα δεδομένων

**D1: Περιλήψεις και Απήχηση Δημοσιεύσεων.** Χρησιμοποιήσαμε μια μεγάλη διεπιστημονική συλλογή δημοσιεύσεων, τις οποίες συλλέξαμε μέσω των αναγνωριστικών ψηφιακών αντικειμένων (Digital Object Identifier - DOI) και οι οποίες καταγράφονταν στο σύνολο δεδομένων ανοιχτών αναφορών OpenCitations COCI. Επιπλέον συλλέξαμε τις περιλήψεις και τους τίτλους των δημοσιεύσεων αυτών από το Open Academic Graph [73, 69] και το Crossref API (βλ. και Ενότητα 5.1.1), κρατώντας τελικά από το αρχικό σύνολο τις δημοσιεύσεις για τις οποίες ήταν διαθέσιμη η περίληψη. Στη συνέχεια «καθαρίσαμε» τα δεδομένα που συλλέχθηκαν, αφαιρώντας τις δημοσιεύσεις που περιείχαν ετικέτες XML (XML tags) στην περίληψη και αυτές που είχαν περίληψη με μέγεθος μικρότερο από τρείς προτάσεις.[1] Αποτέλεσμα όλης της διαδικασίας ήταν η συλλογή $12,534,077$ δημοσιεύσεων και των αναφορών μεταξύ τους (συνολο δεδομένων D1). Τέλος, χρησιμοποιήσαμε αυτό το σύνολο δεδομένων για να υπολογίσουμε τους αριθμούς αναφορών των δημοσιεύσεων και τις βαθμολογίες PageRank και RAM, με χρήση του API της εφαρμογής BIP! (Κεφάλαιο 5).

**D2: Κρίσεις Αναγνωσιμότητας από Ειδικούς.** Συλλέξαμε επιπλέον τις κρίσεις για την αναγνωσιμότητα περιλήψεων από 10 ειδικούς (υποψήφιους διδάκτορες και μεταδιδακτορικούς ερευνητές) του χώρου διαχείρισης δεδομένων και γνώσης (Data and Knowledge Management) μέσω ενός διαδικτυακού ερωτηματολογίου. Οι περιλήψεις που χρησιμοποιήθηκαν ήταν ένα υποσύνολο των δημοσιεύσεων του DBLP που παρέχονται από το AMiner.[2] Για να εγγυηθούμε ότι οι περισσότερες περιλήψεις

[1] Αυτός ο περιορισμός επιβαλλόταν από τη χρήση της βιβλιοθήκης textstat.

[2] https://aminer.org/citation

Πίνακας 6.1: Λίστα όρων που χρησιμοποιήθηκε για την κατασκευή του συνόλου δεδομένων D2.

| | | | |
|---|---|---|---|
| 'database' | 'machine learning' | 'information retrieval' | 'data management' |
| 'cloud computing' | 'data mining' | 'algorithms' | 'classification' |
| 'query processing' | 'networks' | 'indexing' | 'distributed systems' |

Πίνακας 6.2: Τα ερωτήματα του διαδικτυακού ερωτηματολογίου

| | |
|---|---|
| **E1** | «Πόσο καλά γραμμένη είναι η περίληψη;» |
| **E2** | «Υπάρχουν γλωσσικά λάθη στην περίληψη;» |
| **E3** | «Πόσο ξεκάθαρη είναι η συμβολή της συγκεκριμένης εργασίας από την περίληψη;» |

που θα αξιολογούσε ο κάθε ειδικός, θα σχετίζονταν με το αντικείμενο του, χρησιμοποιήσαμε μόνο αυτές που περιείχαν τους όρους που παρουσιάζονται στον Πίνακα 6.1. Κάθε ειδικός παρείχε κρίσεις για ένα μικρό υποσύνολο αυτών των περιλήψεων (από 34 έως 202 περιλήψεις). Αφού ο ειδικός διάβαζε μια περίληψη, απαντούσε σε τρία ερωτήματα σχετικά με την αναγνωσιμότητα της. Τα ερωτήματα αυτά (E1-E3) φαίνονται στον Πίνακα 6.2. Για κάθε ένα από αυτά οι ειδικοί απαντούσαν βάσει μιας κλίμακας 5 πιθανών βαθμολογιών.[3] Κάθε φορά που ένας ειδικός ζητούσε από το σύστημα να αξιολογήσει μια νέα περίληψη, το σύστημα παρουσίαζε είτε μια νέα περίληψη που δεν είχε αξιολογηθεί από κανέναν ειδικό ως τώρα, είτε μια που είχε ήδη αξιολογηθεί από άλλους ειδικούς. Για να εγγυηθούμε μια σημαντική επικάλυψη μεταξύ των περιλήψεων που θα αξιολογούνταν, ακολουθήσαμε την εξής διαδικασία: σε κάθε ειδικό παρουσιαζόταν μια δημοσίευση που δεν είχε αξιολογηθεί από άλλους ως τώρα, μόνο αφού είχε ήδη αξιολογήσει άλλες 10 που είχαν αξιολογηθεί και από άλλους. Το σύνολο δεδομένων D2 που κατασκευάσαμε με αυτόν τον τρόπο είναι ελεύθερα διαθέσιμο στο Zenodo.[4]

### 6.2.2 Μετρικές Αναγνωσιμότητας και Απήχησης.

Στα πειράματά μας εξετάζουμε την αναγνωσιμότητα περιλήψεων βάσει τεσσάρων μετρικών: FRE [25], SMOG [57], Dale-Chall (DC) [92] και Gunning Fog (GF) [32]. Οι δύο πρώτες από αυτές χρησιμοποιούν απλά στατιστικά που προκύπτουν από το κείμενο, όπως το μέσο μήκος των προτάσεων και το μέσο αριθμό συλλαβών ανά λέξη. Αντίθετα οι δύο τελευταίες λαμβάνουν υπόψη και τις «δύσκολες» λέξεις (π.χ., χρησιμοποιώντας κάποιο λεξικό, ή θέτοντας όρια συλλαβών που διαχωρίζουν τις εύκολες από τις δύσκολες λέξεις). Για τη μετρική FRE οι υψηλότερες τιμές αντιστοιχούν σε πιο αναγνώσιμα κείμενα, ενώ για τις υπόλοιπες μετρικές ισχύει το αντίθετο. Όλες οι μετρικές υπολογίστηκαν με χρήση της βιβλιοθήκης textstat της Python.

Επιπλέον υπολογίσαμε την απήχηση των δημοσιεύσεων χρησιμοποιώντας τρεις μετρικές: τον αριθμό αναφορών, τις βαθμολογίες PageRank [59] και τις βαθμολογίες RAM [29]. Ο αριθμός αναφορών είναι το de facto μέτρο που χρησιμοποιείται για την αξιολόγηση της ακαδημαϊκής απόδοσης. Το PageRank μπορεί επιπλέον να διαφοροποιήσει τις αναφορές που λαμβάνει μια δημοσίευση ανάλογα με την αναφέρουσα δημοσίευση, βάσει της αρχής ότι «οι καλές δημοσιεύσεις αναφέρουν άλλες καλές δημοσιεύσεις» (Κεφάλαιο 2). Τέλος οι βαθμολογίες RAM λαμβάνουν υπόψη πόσο πρόσφατα έγιναν οι αναφορές, δίνοντας μεγαλύτερο βάρος στις πιο πρόσφατες. Με αυτό το τρόπο με-

[3]Για κάθε ερώτημα, οι ακραίες τιμές, 1 και 5, αντιστοιχούσαν στην χαμηλότερη και υψηλότερη βαθμολογία αντίστοιχα.

[4]https://doi.org/10.5281/zenodo.2651009

Πίνακας 6.3: Συσχετίσεις ($\rho$ Spearman) μεταξύ μετρικών αναγνωσιμότητας και απήχησης (οι τιμές FRE έχουν αντιστραφεί για λόγους συνοχής). Ο αστερίσκος (*) δηλώνει στατιστική σημαντικότητα με τιμές p-value $p < 10^{-3}$. Ο διπλός αστερίσκος (**) δηλώνει στατιστική σημαντικότητα με τιμές p-value $p < 10^{-5}$.

| | **FRE** | **SMOG** | **DC** | **GF** |
|---|---|---|---|---|
| **Αριθμός Αναφορών** | $-0.0525^{**}$ | $0.0656^{**}$ | $-0.0013^{**}$ | $0.03800^{**}$ |
| **PageRank** | 0.0001 | $0.0076^{**}$ | $-0.01635^{**}$ | $0.0011^{*}$ |
| **RAM** | $0.1169^{**}$ | $0.1257^{**}$ | $0.0397^{**}$ | $0.0837^{**}$ |

τριάζεται η μεροληψία υπέρ των παλιότερων δημοσιεύσεων. Μπορούμε επομένως να χρησιμοποιήσουμε την τελευταία σαν μετρική της δημοφιλίας (όπως προκύπτει και ως η πιο αποτελεσματική στα αντίστοιχα πειράματα του Κεφαλαίου 4), ενώ τις πρώτες δύο ως μετρικές της συνολικής επιρροής.

## 6.3 Αποτελέσματα και Παρατηρήσεις

### 6.3.1 Απήχηση και Παραδοσιακές Μετρικές Αναγνωσιμότητας

Εξετάζουμε αρχικά την συσχέτιση της αναγνωσιμότητας όπως προκύπτει από τις μετρικές που προαναφέραμε στο σύνολο δεδομένων D1. Μετράμε τη συσχέτιση με χρήση του $\rho$ του Spearman για τα ζεύγη των λιστών ιεράρχησης των δημοσιεύσεων του D1 που προκύπτουν από μετρικές για την απήχηση (αριθμός αναφορών, PageRank, RAM) και τις μετρικές αναγνωσιμότητας (FRE, SMOG, DC, GF). Τα αποτελέσματα παρουσιάζονται στον Πίνακα 6.3. Συνολικά παρατηρούμε πολύ ασθενείς συσχετίσεις μεταξύ αναγνωσιμότητας και απήχησης, κάτι το οποίο είναι σε συμφωνία με παλιότερες εργασίες [27, 88, 49]. Παρατηρούμε ακόμα το εξής ενδιαφέρον σημείο: αν και γενικά η αναγνωσιμότητα και η απήχηση φαίνονται ασυσχέτιστες, ωστόσο συσχετίζονται κάπως περισσότερο (χωρίς όμως να φτάνουν μέχρι τη μέτρια συσχέτιση) όταν εξετάζουμε τη δημοφιλία (RAM). Αυτό υπονοεί ότι οι πιο πρόσφατα αναφερόμενες δημοσιεύσεις (που αναδεικνύονται από τη μέθοδο RAM) φαίνεται κατά μέσο όρο να έχουν λιγότερη αναγνωσιμότητα. Η παρατήρηση ότι με το χρόνο η αναγνωσιμότητα των δημοσιεύσεων πέφτει είναι ένα συμπέρασμα που έχει προκύψει και από προηγούμενες μελέτες [49, 62].

### 6.3.2 Απήχηση και Αναγνωσιμότητα βάσει Ειδικών

Εξετάζουμε στη συνέχεια την αναγνωσιμότητα όπως προκύπτει από τις κρίσεις των ειδικών και τη συσχέτισή της με τις μετρικές απήχησης. Μετράμε τη συσχέτιση με τις μετρικές $\tau$ του Kendall και $\rho$ του Spearman. Συγκεκριμένα, μετράμε τη συσχέτιση μεταξύ των λιστών κατάταξης των δημοσιεύσεων βάσει των μετρικών απήχησης και των λιστών κατάταξης βάσει των μέσων όρων των βαθμολογιών που έδωσαν οι ειδικοί για κάθε ερώτημα (E1-E3) του Πίνακα 6.2. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 6.4. Οι σχετικές τιμές που παρατηρούμε είναι παρόμοιες και για τις δύο μετρικές συσχέτισης. Συνολικά παρατηρούμε πολύ χαμηλές τιμές συσχέτισης, οι οποίες επιπλέον δεν είναι στατιστικά σημαντικές. Ένα συμπέρασμα που μπορεί να αντληθεί από αυτά τα αποτελέσματα είναι ότι η αναγνωσιμότητα δε φαίνεται να παίζει κάποιο σημαντικό ρόλο στο κατά πόσο μια δημοσίευση τελικά θα λάβει αναφορές. Τα απο-

Πίνακας 6.4: Συσχετίσεις των κρίσεων των ειδικών με την απήχηση των δημοσιεύσεων.

| | **Spearman's $\rho$** | | | **Kendall's $\tau$** | | |
|---|---|---|---|---|---|---|
| | **Αναφορές** | **PR** | **RAM** | **Αναφορές** | **PR** | **RAM** |
| **E1** | 0.1925 | 0.1896 | 0.2242 | 0.1358 | 0.1286 | 0.1539 |
| **E2** | 0.1827 | 0.1433 | 0.1963 | 0.1273 | 0.0946 | 0.1366 |
| **E3** | 0.162 | 0.1285 | 0.2192 | 0.1139 | 0.0878 | 0.1526 |

τελέσματά μας δείχνουν ότι αυτό ισχύει τόσο εάν εξετάσουμε παραδοσιακές μετρικές αναγνωσιμότητας, όσο και αν εξετάσουμε τις κρίσεις των ειδικών. Οι παρατηρήσεις μας αυτές, σε συνδυασμό με τα συμπεράσματα του [88] μπορούν να θεωρηθούν ως αντεπιχείρημα σε ισχυρισμούς ότι οι απλές περιλήψεις συσχετίζονται με τις αναφορές των δημοσιεύσεων, όπως γίνεται στο [50].

## 6.4 Σύνοψη

Σε αυτό το κεφάλαιο εξετάσαμε κατά πόσο η αναγνωσιμότητα των περιλήψεων επιστημονικών δημοσιεύσεων συσχετίζεται με διάφορες μετρικές της απήχησης τους. Εξετάσαμε τόσο την περίπτωση που η αναγνωσιμότητα μετριέται με παραδοσιακές μετρικές που την ποσοτικοποιούν με βάση επιφανειακά κειμενικά χαρακτηριστικά, όσο και την περίπτωση που καταγράφουμε τις κρίσεις ειδικών. Σε κάθε περίπτωση τα αποτελέσματά μας δείχνουν ότι δεν υπάρχει συσχέτιση μεταξύ της απήχησης και της αναγνωσιμότητας των δημοσιεύσεων.

# Κεφάλαιο 7

# Συμπεράσματα και Μελλοντικές Εργασίες

Στην παρούσα διατριβή μελετήσαμε μεθόδους κατάταξης επιστημονικών δημοσιεύσεων με βάση την απήχησή τους και εξετάσαμε τα προβλήματα κατάταξης με βάση δύο είδη απήχησης, την επιρροή και τη δημοφιλία. Παρουσιάσαμε μια ευρεία κατηγοριοποίηση της τρέχουσας βιβλιογραφίας που αφορά μεθόδους κατάταξης δημοσιεύσεων, εξετάζοντας τους διαφορετικούς μηχανισμούς που έχουν χρησιμοποιηθεί. Επιπλέον, εξετάσαμε τις μεθοδολογίες πειραματικής αξιολόγησης που εντοπίζονται στην παρούσα βιβλιογραφία. Ακόμα, επιλέξαμε ένα σύνολο μεθόδων ώστε να καλύπτουμε όλες τις κατηγορίες προσεγγίσεων που καταγράψαμε και αξιολογήσαμε πειραματικά τις μεθόδους αυτές, προτείνοντας ένα συγκεκριμένο πλαίσιο αξιολόγησης με βάση το οποίο μπορούμε να αποφασίσουμε ποιες είναι αποτελεσματικότερες στην παραγωγή κατατάξεων με βάση τη δημοφιλία και την επιρροή. Ακόμα, προτείναμεμια νέα μέθοδο που πετυχαίνει αποτελεσματικότερη παραγωγή κατατάξεων με βάση τη δημοφιλία. Επιπλέον, εφαρμόσαμε το πλαίσιο αξιολόγησής μεθόδων που προτείναμε για να επιλέξουμε και να παραμετροποιήσουμε αποτελεσματικές μεθόδους κατάταξης που εφαρμόζονται σε πραγματικά συστήματα αναζήτησης επιστημονικών δημοσιεύσεων. Τέλος, πραγματοποιήσαμε μια μελέτη της συσχέτισης της απήχησης των επιστημονικών δημοσιεύσεων με την αναγνωσιμότητα των περιλήψεων τους.

Στη συνέχεια συνοψίζουμε πιο λεπτομερώς τις συνεισφορές της εργασίας μας και παρουσιάζουμε κατευθύνσεις που προσφέρονται για μελλοντική έρευνα.

## 7.1 Σύνοψη

Αρχικά ορίσαμε τυπικά τα προβλήματα της κατάταξης με βάση την επιρροή και τη δημοφιλία. Στη συνέχεια εξετάσαμε ενδελεχώς τις μεθόδους κατάταξης δημοσιεύσεων που έχουν προταθεί στη βιβλιογραφία. Αυτές ταξινομήθηκαν σε κατηγορίες με βάση τις διάφορες προσεγγίσεις που χρησιμοποιούν. Συνοπτικά διακρίνουμε απλές παραλλαγές του PageRank, μεθόδους που χρησιμοποιούν χρονικούς παράγοντες είτε στους πίνακες γειτνίασης, είτε στους πίνακες μετάβασης, είτε στις πιθανότητες επιλογής του PageRank. Ακόμα, διακρίνουμε μεθόδους που χρησιμοποιούν μεταδεδομένα, όπως ο χώρος δημοσίευσης και οι συγγραφείς. Επιπλέον, εντοπίζουμε μεθόδους που κάνουν χρήση πολλαπλών δικτύων καθώς και μεθόδους που συνδυάζουν όλες τις παραπάνω προσεγγίσεις, συναθροίζοντας τα διάφορα αποτελέσματα. Τέλος εντοπίζουμε ένα μικρό σύνολο μεθόδων οι οποίες δεν εμπίπτουν σε καμία από τις παραπάνω κατηγορίες.

Παράλληλα εξετάσαμε τις προσεγγίσεις για την αξιολόγηση των μεθόδων που ακολουθούνται στη βιβλιογραφία. Συνοπτικά εντοπίσαμε δύο κατηγορίες μεθοδολογιών: αυτές που αξιολογούν την αποτελεσματικότητα κατατάξεων με κάποια αντικειμενικά κριτήρια και τις μεθοδολογίες που δε χρησιμοποιούν κριτήρια σχετικά με τις παραγόμενες κατατάξεις. Στην πρώτη κατηγορία εντοπίζουμε τις μεθοδολογίες που χρησιμοποιούν υπόβαθρα αληθείας με βραβευμένες δημοσιεύσεις, μεθοδολογίες που βασίζονται σε κρίσεις χρηστών, μεθοδολογίες που βασίζονται σε παρακράτηση δεδομένων. Στη δεύτερη κατηγορία εντοπίζουμε τις περιγραφικές αξιολογήσεις και τις αξιολογήσεις που αφορούν ειδικά χαρακτηριστικά των μεθόδων όπως η ταχύτητα σύγκλισης, ή σταθερότητα σε μεταβολές παραμέτρων κλπ.

Έπειτα επιλέξαμε ένα σύνολο μεθόδων έτσι ώστε να καλύπτονται όλες οι προσεγγίσεις της βιβλιογραφίας και προτείναμε ένα πλαίσιο αξιολόγησης που βασίζεται σε παρακράτηση δεδομένων. Βάσει αυτού του πλαισίου αξιολογήσαμε την αποτελεσματικότητα των μεθόδων της βιβλιογραφίας στην κατάταξη με βάση τη δημοφιλία και την επιρροή. Τα αποτελέσματα της ανάλυσής μας έδειξαν ότι με βάση την επιρροή οι πιο αποτελεσματικές μέθοδοι είναι οι PageRank και CiteRank, ενώ με βάση τη δημοφιλία υπερισχύουν οι RAM και ECM. Επιπλέον η ανάλυσή μας έδειξε ότι το πρόβλημα κατάταξης με βάση την επιρροή καλύπτεται επαρκώς από τρέχουσες τεχνολογίες, ενώ στο πρόβλημα της κατάταξης με βάση την δημοφιλία υπάρχουν περιθώρια βελτίωσης.

Βασισμένοι στα παραπάνω αποτελέσματα χρησιμοποιήσαμε αποτελεσματικές μεθόδους κατάταξης σε πραγματικά συστήματα αναζήτησης δημοσιεύσεων, το mirPub v2 και το BIP! Finder. Το πρώτο σύστημα έχει αναπτυχθεί για αναζήτηση βιβλιογραφίας στον τομέα της μοριακής βιολογίας, συγκεκριμένα βιβλιογραφίας που σχετίζεται με τα miRNAs. Οι μέθοδοι κατάταξης που χρησιμοποιεί το σύστημα παραμετροποιήθηκαν έτσι ώστε να είναι αποτελεσματικές στην παραγωγή κατατάξεων με βάση την επιρροή. Από την άλλη το σύστημα BIP! είναι μια ακαδημαϊκή μηχανή αναζήτησης γενικού σκοπού. Η ιδιαιτερότητά του είναι ότι επιτρέπει στο χρήστη να επιλέγει το κριτήριο με βάση το οποίο ιεραρχούνται τα αποτελέσματα, ενώ προσφέρει τη δυνατότητα πολλών οπτικοποιήσεων που αφορούν την απήχηση και άλλα χαρακτηριστικά των δημοσιεύσεων. Επιπλέον, δεδομένου του περιθωρίου βελτίωσης που εντοπίσαμε στο πρόβλημα κατάταξης με βάση την δημοφιλία, προτείναμε μια νέα μέθοδο κατάταξης. Η μέθοδος αυτή βασίζεται σε ένα μηχανισμό πρόσφατης προσοχής, ο οποίος αποδίδει αξία σε δημοσιεύσεις εφόσον έχουν λάβει πολλές αναφορές σε ένα πρόσφατο χρονικό παράθυρο. Πραγματοποιήσαμε εξαντλητικά πειράματα που αφενός ανέδειξαν τη σημασία αυτού του μηχανισμού και αφετέρου έδειξαν ότι η μέθοδος υπερισχύει έναντι άλλων τεχνολογιών αιχμής στη βιβλιογραφία.

Τέλος, θεωρώντας ότι η αξία των επιστημονικών δημοσιεύσεων μπορεί να εξεταστεί μέσω πολλών διαφορετικών χαρακτηριστικών, πραγματοποιήσαμε μια μελέτη της συσχέτισης της απήχησης των δημοσιεύσεων με την αναγνωσιμότητά τους. Η ανάλυσή μας έδειξε ότι τα δύο αυτά χαρακτηριστικά είναι ασυσχέτιστα, σε συμφωνία με προηγούμενες εργασίες.

## 7.2 Μελλοντικές Εργασίες

Κατά τη διάρκεια των παραπάνω εργασιών πέρα από τα συμπεράσματα που αντλήθηκαν, εντοπίστηκαν και μια σειρά από ανοιχτά ζητήματα που προσφέρονται για μελλοντικές εργασίες.

- **Χρήση μεταδεδομένων.** Αν και είδαμε στο Κεφάλαιο 3 ότι μια σειρά μεθόδων κάνουν χρήση μεταδεδομένων (π.χ., συγγραφέων, περιοδικών κλπ), ωστόσο στο Κεφάλαιο 4 είδαμε ότι αυτές οι μέθοδοι δεν παράγουν αποτελεσματικές κατατάξεις με βάση τη δημοφιλία και την επιρροή. Αυτό δε σημαίνει υποχρεωτικά ότι είναι άνευ νοήματος η χρήση μεταδεδομένων. Αντίθετα, θεωρούμε ότι υπάρχει χώρος για ορθότερη χρήση τους, για παράδειγμα μέσω τεχνικών μηχανικής μάθησης (Machine Learning - ML). Για παράδειγμα στις εργασίες [16, 82, 64] οι συγγραφείς προτείνουν οι μελλοντικές εργασίες να συνδυάζουν την ανάλυση του γράφου αναφορών με τέτοιες τεχνικές, με σκοπό π.χ. να εκπαιδευτούν κατάλληλα βάρη που μπορούν να ενσωματωθούν στις εξισώσεις των μεθόδων. Μια πιθανή προσέγγιση σε αυτή τη κατεύθυνση είναι η χρήση τεχνικών πρόβλεψης αναφορών [13, 86, 81]. Οι τεχνικές αυτές χρησιμοποιούν μηχανική μάθηση πάνω στα μεταδεδομένα για να προβλέψουν μελλοντικούς αριθμούς αναφορών και έτσι τα αποτελέσματά τους θα μπορούσαν να ενσωματώνονται σε μελλοντικές μεθόδους κατάταξης. Επιπλέον, νέα είδη μεταδεδομένων, οι λεγόμενες εναλλακτικές μετρικές (Altmetrics) [63, 61, 23] είναι σήμερα διαθέσιμα και αφορούν στατιστικά χρήσης, όπως το πόσες φορές έχουν ανοίξει χρήστες τη σελίδα ενός άρθρου, ή πόσοι χρήστες το έχουν κατεβάσει. Τέτοια μεταδεδομένα δεν έχουν χρησιμοποιηθεί ακόμα σε μεθόδους κατάταξης με βάση την απήχηση και θα μπορούσαν να ενσωματωθούν επίσης σε μελλοντικές μεθόδους.

- **Αξιολόγηση και Μετρικές.** Ανοιχτό ζήτημα αποτελεί επιπλέον η προτυποποίηση μεθοδολογιών για την αξιολόγηση μεθόδων κατάταξης. Στη παρούσα διατριβή διατυπώσαμε τυπικούς ορισμούς για την κατάταξη με βάση την δημοφιλία και την επιρροή, δύο είδη απήχησης που ως τώρα στη βιβλιογραφία συχνά δε διαχωρίζονται ή συγχέονται. Επιπλέον παρουσιάσαμε ένα πλαίσιο για την αξιολόγηση μεθόδων κατάταξης όσον αφορά την αποτελεσματικότητά τους να ιεραρχούν δημοσιεύσεις με βάση την απήχηση. Ωστόσο η παραγωγή κατατάξεων με βάση άλλες ενδιαφέρουσες ιδιότητες των δημοσιεύσεων (όπως λ.χ., το πόσο καινοτόμες είναι οι ιδέες τους) παραμένει ένα ζήτημα προς διερεύνηση. Φυσικά κάθε τέτοια διερεύνηση είναι άρρηκτα συνδεδεμένη με την ανάγκη για τυπικό ορισμό αυτών των ιδιοτήτων, καθώς και πλαισίων αξιολόγησης με βάση αυτές.

- **Κατασκευή συνόλων δεδομένων.** Ένα σοβαρό μειονέκτημα προηγούμενων εργασιών είναι η έλλειψη πειραματικών αξιολογήσεων σε πολλαπλά σύνολα δεδομένων. Πολλές προηγούμενες εργασίες χρησιμοποιούν λίγα, αν όχι ένα μοναδικό σύνολο δεδομένων. Στη παρούσα διατριβή εκτελέσαμε πειράματα σε τέσσερα σύνολα δεδομένων ώστε να βγάλουμε όσο το δυνατόν πιο γενικεύσιμα συμπεράσματα. Επιπλέον τα πειράματα που δημοσιεύονται θα έπρεπε να είναι αναπαραγώγιμα (Reproducible), σε επιπλέον σύνολα δεδομένων διαφορετικών μεγεθών. Επιπλέον το κίνημα προς την ανοιχτή επιστήμη [21] ορίζει την αναπαραγωγιμότητα της επιστημονικής έρευνας ως βασική αρχή της επιστημονικής διαδικασίας. Συνεπώς πρέπει να κατασκευαστούν καινούρια, μεγαλύτερα σύνολα δεδομένων που θα γίνουν διαθέσιμα στην επιστημονική κοινότητα για περαιτέρω έρευνα.

- **Αντιμετώπιση Κακόβουλων Συμπεριφορών.** Η χρήση διαφόρων ποσοτικών μέτρων για την αξιολόγηση της επιστημονικής έρευνας και των ερευνητών έχει ως αποτέλεσμα και την εμφάνιση κακόβουλων πρακτικών στους ακαδημα-

ϊκούς κύκλους. Συγκεκριμένα είναι γνωστή η κατάχρηση αυτο-αναφορών, ή αμοιβαίων αναφορών από κάποια μέλη της ακαδημαϊκής κοινότητας. Αυτή η κατάχρηση έχει οδηγήσει μέχρι και στην πρόταση μετρικών αυτο-αναφορών όπως ο S-index [24], μια μετρική εμπνευσμένη από τον παραδοσιακό h-index. Αυτά τα ζητήματα έχουν αναδειχτεί σε κάποιο βαθμό από μερικές εργασίες που εξετάσαμε στη διατριβή [18, 87]. Αν και κάποιες μέθοδοι κατάταξης της βιβλιογραφίας έχουν ως αφετηρία την αντιμετώπιση των αυτο-αναφορών (π.χ., [3]), ή των αμοιβαίων αναφορών (π.χ., [52]), υπάρχουν ακόμα ανοιχτά ζητήματα. Για παράδειγμα η μελέτη του πόσο σταθερά είναι τα αποτελέσματα των μεθόδων κατάταξης απέναντι σε τέτοιες πρακτικές, ή πώς μπορεί να γίνει διαχωρισμός αυτο-αναφορών που είναι έγκυρες, έναντι όσων γίνονται κακόβουλα.

- **Βελτιωμένες Κατατάξεις Δημοφιλίας.** Στη διατριβή προτείναμε μια νέα μέθοδο που παράγει καλύτερες κατατάξεις με βάση τη δημοφιλία σε σχέση με ανταγωνίστριες μεθόδους της βιβλιογραφίας. Το καινοτόμο στοιχείο της μεθόδου είναι ότι διαχωρίζει τις δημοσιεύσεις που εκδόθηκαν το ίδιο έτος με βάση τον αριθμό αναφορών που έχουν λάβει, σε αντίθεση με άλλες μεθόδους που χρησιμοποιούν χρονικούς παράγοντες και ομαδοποιούν ενιαία *όλες* τις δημοσιεύσεις που εκδόθηκαν το ίδιο έτος, αποδίδοντάς τους το ίδιο βάρος. Ωστόσο ένα μεγάλο πλήθος νέων δημοσιεύσεων έχουν μηδενικό αριθμό αναφορών. Ανάμεσα τους κάποιες πράγματι θα λάβουν αναφορές και άλλες όχι. Η ανάπτυξη μηχανισμών που μπορούν να διαχωρίσουν αυτές τις δύο περιπτώσεις και η εισαγωγή τους σε επεκτάσεις της μεθόδου μας μπορεί να δώσει νέα βελτιωμένα αποτελέσματα κατατάξεων με βάση τη δημοφιλία.

- **Αποτελεσματικότερες Μετρήσεις Αναγνωσιμότητας.** Οι μετρικές αναγνωσιμότητας που χρησιμοποιήσαμε στο Κεφάλαιο 6 βασίζονται σε επιφανειακά κειμενικά χαρακτηριστικά, όπως η δυσκολία λέξεων και το μήκος των προτάσεων. Ως προς τέτοια χαρακτηριστικά είναι αρκετά αναμενόμενο όλες οι επιστημονικές δημοσιεύσεις να χαρακτηρίζονται ως «δύσκολες», αφού η επιστήμη παράγεται εκ των πραγμάτων από ανθρώπους υψηλής τεχνικής κατάρτισης. Συνεπώς υπάρχει περιθώριο ανάπτυξης εξειδικευμένων μετρικών αναγνωσιμότητας που θα αφορούν ένα καταρτισμένο κοινό. Επιπλέον, μπορούν να χρησιμοποιηθούν πιο αντικειμενικά ποσοτικοποιήσιμες μέθοδοι για να μετρηθεί η αναγνωσιμότητα των δημοσιεύσεων όπως δίνεται υποκειμενικά από ειδικούς, π.χ., μέσω της χρονομέτρησης της διάρκειας ανάγνωσης των περιλήψεων. Συμπεράσματα από τέτοιες μελέτες προσφέρονται επιπλέον και για μια επανεκτίμηση της συσχέτισης της αναγνωσιμότητας με την απήχηση των δημοσιεύσεων.

- **Επεκτάσεις Μηχανών Αναζήτησης.** Στο Κεφάλαιο 5 περιγράψαμε ακαδημαϊκές μηχανές αναζήτησης που αναπτύξαμε. Ειδικά η μηχανή αναζήτησης BIP! Finder δίνει στους χρήστες τη δυνατότητα να συγκρίνουν δημοσιεύσεις ως προς διάφορα χαρακτηριστικά σε διαγράμματα ραντάρ. Μια μελλοντική κατεύθυνση αφορά τον εμπλουτισμό των διαστάσεων με βάση τις οποίες συγκρίνονται οι δημοσιεύσεις, με πιθανούς νέους άξονες για την καινοτομία, την παρουσία άρθρων στα μέσα κοινωνικής δικτύωσης, κλπ. Επιπλέον χρήσιμες υπάρχει περιθώριο για την ανάπτυξη χρήσιμων επεκτάσεων που μπορούν να κάνουν χρήση της απήχησης των δημοσιεύσεων, όπως π.χ., η ανάπτυξη ενός συστήματος προτάσεων (Recommender) το οποίο μπορεί να δίνει έμφαση σε σχετικές εργασίες

με βάση την απήχησή τους.

Συμπερασματικά, θεωρούμε ότι υπάρχει ένα πλήθος από ενδιαφέροντα θέματα για περαιτέρω έρευνα πάνω στο ζήτημα της κατάταξης δημοσιεύσεων με βάση την απήχηση, του εντοπισμού άλλων ποιοτικών χαρακτηριστικών των δημοσιεύσεων και της συσχέτισης τους με την απήχηση. Ελπίζουμε η παρούσα διατριβή να αποτελέσει ένα εφαλτήριο για επιπλέον έρευνα σε αυτό το πεδίο.

# Bibliography


[1] X. Bai, J. Hou, H. Du, X. Kong, and F. Xia. Evaluating the impact of articles with geographical distances between institutions. In *Proceedings of the 26th international conference on world wide web companion*, pages 1243–1244. International World Wide Web Conferences Steering Committee, 2017.

[2] X. Bai, H. Liu, F. Zhang, Z. Ning, X. Kong, I. Lee, and F. Xia. An overview on evaluating and predicting scholarly article impact. *Information*, 8(3):73, 2017.

[3] X. Bai, F. Xia, I. Lee, J. Zhang, and Z. Ning. Identifying anomalous citations for objective evaluation of scholarly article impact. *PloS one*, 11(9):e0162364, 2016.

[4] A.-L. Barabási et al. *Network science*. Cambridge university press, 2016.

[5] C. T. Bergstrom, J. D. West, and M. A. Wiseman. The eigenfactor$^{\text{TM}}$ metrics. *The Journal of Neuroscience*, 28(45):11433–11434, 2008.

[6] E. V. Bernstam, J. R. Herskovic, Y. Aphinyanaphongs, C. F. Aliferis, M. G. Sriram, and W. R. Hersh. Using citation data to improve retrieval from medline. *Journal of the American Medical Informatics Association*, 13(1):96–105, 2006.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[8] J. Bollen, H. Van de Sompel, A. Hagberg, and R. Chute. A principal component analysis of 39 scientific impact measures. *PloS one*, 4(6):e6022, 2009.

[9] L. Bornmann and R. Mutz. Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11):2215–2222, 2015.

[10] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117, 1998.

[11] M. Brzezinski. Power laws in citation distributions: Evidence from scopus. *Scientometrics*, 103(1):213–228, 2015.

[12] R. Carthew. Gene regulation by microrнas. *Curr. Opin. Genet. Dev.*, 16(2):203–208, 2006.

[13] C. Castillo, D. Donato, and A. Gionis. Estimating number of citations using author reputation. In *International Symposium on String Processing and Information Retrieval*, pages 107–117. Springer, 2007.

[14] S. Chang, S. Go, Y. Wu, Y. Lee, C. Lai, S. Yu, C. Chen, H. Chen, M. Tsai, M. Yeh, and S. Lin. An ensemble of ranking strategies for static rank prediction in a large heterogeneous graph. *WSDM Cup*, 2016.

[15] P. Chen, H. Xie, S. Maslov, and S. Redner. Finding scientific gems with google's pagerank algorithm. *Journal of Informetrics*, 1(1):8–15, 2007.

[16] H. Chin-Chi, C. Kuan-Hou, F. Ming-Han, W. Yueh-Hua, C. Huan-Yuan, Y. Sz-Han, C. Chun-Wei, T. Ming-Feng, Y. Mi-Yen, and L. Shou-De. Time-aware weighted pagerank for paper ranking in academic graphs. *WSDM Cup*, 2016.

[17] V. P. Diodato and P. Gellatly. *Dictionary of Bibliometrics (Haworth Library and Information Science)*. Routledge, 1994.

[18] M. Dunaiski and W. Visser. Comparing paper ranking algorithms. In *Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference*, pages 21–30. ACM, 2012.

[19] Y.-H. Eom and S. Fortunato. Characterizing and modeling citation dynamics. *PloS one*, 6(9):e24926, 2011.

[20] J. D. Evans. *Straightforward statistics for the behavioral sciences*. Brooks/Cole, 1996.

[21] A. Farnham, C. Kurz, M. A. Öztürk, M. Solbiati, O. Myllyntaus, J. Meekes, T. M. Pham, C. Paz, M. Langiewicz, S. Andrews, et al. Early career researchers want open science. *Genome biology*, 18(1):221, 2017.

[22] M. Feng, K. Chan, H. Chen, M. Tsai, M. Yeh, and S. Lin. An efficient solution to reinforce paper ranking using author/venue/citation information-the winner's solution for wsdm cup 2016. *WSDM Cup*, 2016.

[23] M. Fenner. Altmetrics and other novel measures for scientific impact. In *Opening science*, pages 179–189. Springer, 2014.

[24] J. Flatt, A. Blasimme, and E. Vayena. Improving the measurement of scientific success by reporting a self-citation index. *Publications*, 5(3):20, 2017.

[25] R. Flesch. A New Readability Yardstick. *Journal of applied psychology*, 32(3):221, 1948.

[26] E. Garfield. The history and meaning of the journal impact factor. *Jama*, 295(1):90–93, 2006.

[27] A. Gazni. Are the Abstracts of High Impact Articles more Readable? Investigating the Evidence from Top Research Institutions in the World. *J. Information Science*, 37(3):273–281, 2011.

[28] R. Ghosh, T.-T. Kuo, C.-N. Hsu, S.-D. Lin, and K. Lerman. Time-aware ranking in dynamic citation networks. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 373–380. IEEE, 2011.

[29] R. Ghosh, T.-T. Kuo, C.-N. Hsu, S.-D. Lin, and K. Lerman. Time-aware ranking in dynamic citation networks. In *Data Mining Workshops (ICDMW)*, pages 373–380, 2011.

[30] S. Griffiths-Jones, R. Grocock, S. Van Dongen, A. Bateman, and A. Enright. mirbase: microrna sequences, targets and gene nomenclature. *Nucleic Acids Res.*, 34(suppl 1):D140–D144, 2006.

[31] P. Groth and T. Gurney. Studying scientific discourse on the web using bibliometrics: A chemistry blogging case study. In *WebSci*, 2010.

[32] R. Gunning. *The Technique of Clear Writing*. McGraw-Hill, New York, 1952.

[33] J. Hartley, E. Sotto, and J. Pennebaker. Style and Substance in Psychology: Are Influential Articles more Readable than less Influential Ones? *Social Studies of Science*, 32(2):321–334, 2002.

[34] T. H. Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *Knowledge and Data Engineering, IEEE Transactions on*, 15(4):784–796, 2003.

[35] D. Herrmannova and P. Knoth. Simple yet effective methods for large-scale scholarly publication ranking. *arXiv preprint arXiv:1611.05222*, 2016.

[36] B. I. Hutchins, X. Yuan, J. M. Anderson, and G. M. Santangelo. Relative citation ratio (rcr): A new metric that uses citation rates to measure influence at the article level. *PLoS biology*, 14(9):e1002541, 2016.

[37] W.-S. Hwang, S.-M. Chae, S.-W. Kim, and G. Woo. Yet another paper ranking algorithm advocating recent publications. In *Proceedings of the 19th international conference on World wide web*, pages 1117–1118. ACM, 2010.

[38] J. P. Ioannidis. Why most published research findings are false. *PLoS Med*, 2(8):e124, 2005.

[39] G. Jeh and J. Widom. Scaling personalized web search. In *WWW*, pages 271–279. ACM, 2003.

[40] X. Jiang, X. Sun, and H. Zhuge. Towards an effective and unbiased ranking of scientific literature through mutual reinforcement. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 714–723. ACM, 2012.

[41] M. G. Kendall. *Rank correlation methods*. Hafner Publishing Co., 1955.

[42] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[43] D. F. Klosik and S. Bornholdt. The citation wake of publications detects nobel laureates' papers. *PloS one*, 9(12):e113184, 2014.

[44] V. Koukis, C. Venetsanopoulos, and N. Koziris. ˜ okeanos: Building a cloud, cluster by cluster. *IEEE internet computing*, 17(3):67–71, 2013.

[45] A. Kozomara and S. Griffiths-Jones. mirbase: annotating high confidence micrornas using deep sequencing data. *Nucleic acids research*, page gkt1181, 2013.

[46] M. Krapivin and M. Marchese. Focused page rank in scientific papers ranking. In *International Conference on Asian Digital Libraries*, pages 144–153. Springer, 2008.

[47] A. N. Langville and C. D. Meyer. *Google's PageRank and beyond: The science of search engine rankings*. Princeton University Press, 2011.

[48] P. O. Larsen and M. von Ins. The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics*, 84(3):575–603, 2010.

[49] L. Lei and S. Yan. Readability and Citations in Information Science: Evidence from Abstracts and Articles of Four Journals (2003-2012). *Scientometrics*, 108(3):1155–1169, 2016.

[50] A. Letchford, T. Preis, and H. S. Moat. The Advantage of Simple Paper Abstracts. *Journal of Informetrics*, 10(1):1–8, 2016.

[51] H. Liao, M. S. Mariani, M. Medo, Y.-C. Zhang, and M.-Y. Zhou. Ranking in evolving complex networks. *Physics Reports*, 689:1–54, 2017.

[52] Z. Liu, H. Huang, X. Wei, and X. Mao. Tri-rank: An authority ranking framework in heterogeneous academic networks by mutual reinforce. In *2014 IEEE 26th International Conference on Tools with Artificial Intelligence*, pages 493–500. IEEE, 2014.

[53] D. Luo, C. Gong, R. Hu, L. Duan, and S. Ma. Ensemble enabled weighted pagerank. *arXiv preprint arXiv:1604.05462*, 2016.

[54] N. Ma, J. Guan, and Y. Zhao. Bringing pagerank to the citation analysis. *Information Processing & Management*, 44(2):800–810, 2008.

[55] S. Ma, C. Gong, R. Hu, D. Luo, C. Hu, and J. Huai. Query independent scholarly article ranking. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, pages 953–964. IEEE, 2018.

[56] M. S. Mariani, M. Medo, and Y.-C. Zhang. Identification of milestone papers through time-balanced network centrality. *Journal of Informetrics*, 10(4):1207–1223, 2016.

[57] G. H. Mc Laughlin. Smog Grading-a New Readability Formula. *Journal of reading*, 12(8):639–646, 1969.

[58] Z. Nie, Y. Zhang, J.-R. Wen, and W.-Y. Ma. Object-level ranking: bringing order to web objects. In *Proceedings of the 14th international conference on World Wide Web*, pages 567–574. ACM, 2005.

[59] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web., 1999.

[60] G. D. Paparo and M. Martin-Delgado. Google in a quantum network. *Scientific reports*, 2:444, 2012.

[61] H. Piwowar. Introduction altmetrics: What, why and where? *Bulletin of the American Society for Information Science and Technology*, 39(4):8–9, 2013.

[62] P. Plavén-Sigray, G. J. Matheson, B. C. Schiffler, and W. H. Thompson. The Readability of Scientific Texts is Decreasing over Time. *Elife*, page e27725, 2017.

[63] J. Priem, D. Taraborelli, P. Groth, and C. Neylon. Altmetrics: A manifesto. 2010.

[64] S. Ribas, A. Ueda, R. L. Santos, B. Ribeiro-Neto, and N. Ziviani. Simplified relative citation ratio for static paper ranking: Ufmg/latin at wsdm cup 2016. *arXiv preprint arXiv:1603.01336*, 2016.

[65] D. Sarewitz. The pressure to publish pushes down quality. *Nature*, 533(7602):147–147, 2016.

[66] H. Sayyadi and L. Getoor. Futurerank: Ranking scientific articles by predicting their future pagerank. In *SDM*, pages 533–544. SIAM, 2009.

[67] A. Sidiropoulos and Y. Manolopoulos. Generalized comparison of graph-based ranking algorithms for publications and authors. *Journal of Systems and Software*, 79(12):1679–1700, 2006.

[68] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-j. P. Hsu, and K. Wang. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246. ACM, 2015.

[69] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. P. Hsu, and K. Wang. An overview of microsoft academic service (mas) and applications. In *WWW (Companion Volume)*, pages 243–246, 2015.

[70] D. R. Smith. A 30-year citation analysis of bibliometric trends at the archives of environmental health, 1975–2004. *Archives of environmental & occupational health*, 64(sup1):43–54, 2009.

[71] C. Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.

[72] C. Su, Y. Pan, Y. Zhen, Z. Ma, J. Yuan, H. Guo, Z. Yu, C. Ma, and Y. Wu. Prestigerank: A new evaluation method for papers and journals. *Journal of Informetrics*, 5(1):1–13, 2011.

[73] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: Extraction and mining of academic social networks. In *SIGKDD*, pages 990–998, 2008.

[74] G. Vaccario, M. Medo, N. Wider, and M. S. Mariani. Quantifying and suppressing ranking bias in a large citation network. *Journal of informetrics*, 11(3):766–782, 2017.

[75] T. Vergoulis, I. Kanellos, N. Kostoulas, G. Georgakilas, T. Sellis, A. Hatzigeorgiou, and T. Dalamagas. mirpub: a database for searching microrna publications. *Bioinformatics*, 31(9):1502–1504, 2015.

[76] I. Vlachos, M. Paraskevopoulou, D. Karagkouni, G. Georgakilas, T. Vergoulis, I. Kanellos, I. Anastasopoulos, S. Maniou, K. Karathanou, D. Kalfakakou, et al. Diana-tarbase v7. 0: indexing more than half a million experimentally supported mirna: mrna interactions. *Nucleic Acids Res.*, 43(D1):D153–D159, 2015.

[77] A. D. Wade, K. Wang, Y. Sun, and A. Gulli. Wsdm cup 2016: Entity ranking challenge. In *Proceedings of the ninth ACM international conference on web search and data mining*, pages 593–594. ACM, 2016.

[78] D. Walker, H. Xie, K.-K. Yan, and S. Maslov. Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(06):P06010, 2007.

[79] D. Wang, C. Song, and A. Barabási. Quantifying long-term scientific impact. *CoRR*, abs/1306.3293, 2013.

[80] Y. Wang, Y. Tong, and M. Zeng. Ranking scientific articles by exploiting citations, authors, journals, and time information. In *AAAI*, 2013.

[81] L. Weihs and O. Etzioni. Learning to predict citation-based impact measures. In *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries*, pages 49–58. IEEE Press, 2017.

[82] I. Wesley-Smith, C. T. Bergstrom, and J. D. West. Static ranking of scholarly papers using article-level eigenfactor (alef). *arXiv preprint arXiv:1606.08534*, 2016.

[83] B. Xie, Q. Ding, H. Han, and D. Wu. mircancer: a microrna–cancer association database constructed by text mining on literature. *Bioinformatics*, page btt014, 2013.

[84] E. Yan and Y. Ding. Weighted citation: An indicator of an article's prestige. *Journal of the American Society for Information Science and Technology*, 61:1635–1643, August 2010.

[85] E. Yan, Y. Ding, and C. R. Sugimoto. P-rank: An indicator measuring prestige in heterogeneous scholarly networks. *Journal of the American Society for Information Science and Technology*, 62(3):467–477, 2011.

[86] R. Yan, J. Tang, X. Liu, D. Shan, and X. Li. Citation count prediction: learning to estimate future citations for literature. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1247–1252. ACM, 2011.

[87] L. Yao, T. Wei, A. Zeng, Y. Fan, and Z. Di. Ranking scientific publications: the effect of nonlinearity. *Scientific reports*, 4, 2014.

[88] A. W. K. Yeung, T. K. Goto, and W. K. Leung. Readability of the 100 Most-Cited Neuroimaging Papers Assessed by Common Readability Formulae. *Frontiers in human neuroscience*, 12:308, 2018.

[89] P. S. Yu, X. Li, and B. Liu. On the temporal dimension of search. In *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, pages 448–449. ACM, 2004.

[90] P. S. Yu, X. Li, and B. Liu. Adding the temporal dimension to search-a case study in publication search. In *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*, pages 543–549. IEEE, 2005.

[91] L. A. Zadeh. Fuzzy sets. *Information and control*, 8(3):338–353, 1965.

[92] M. Zamanian and P. Heydari. Readability of Texts: State of the Art. *Theory & Practice in Language Studies*, 2(1), 2012.

[93] F. Zhang and S. Wu. Ranking scientific papers and venues in heterogeneous academic networks by mutual reinforcement. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, pages 127–130. ACM, 2018.

[94] L. Zhang, S. Volinia, T. Bonome, G. Calin, J. Greshock, N. Yang, C. Liu, A. Giannakakis, P. Alexiou, K. Hasegawa, et al. Genomic and epigenetic alterations deregulate microrna expression in human epithelial ovarian cancer. *P. Natl. A. Sci.-Biol.*, 105(19):7004–7009, 2008.

[95] J. Zhou, A. Zeng, Y. Fan, and Z. Di. Ranking scientific publications with similarity-preferential mechanism. *Scientometrics*, 106(2):805–816, 2016.

| Μετάφραση | Αγγλικός Όρος |
|---|---|
| Αγαπημένα | Bookmarks |
| Αμερικάνικη Ένωση Φυσικών | American Physical Society, APS |
| Αμερικάνικος Οργανισμός Υγείας | National Institute of Health (NIH) |
| Αμοιβαία Αναφορά | Mutual Citaton |
| Αμοιβαία Ενίσχυση | Mutual Reinforcement |
| Αναγνωσιμότητα | Readability |
| Ανάκτηση Πληροφορίας | Information Retrieval |
| Αναπαραγώγιμος | Reproducible |
| Αναφέρομαι | Cite |
| Αναφορά | Reference |
| Ανοικτή Επιστήμη | Open Science |
| Αντικείμενο του Διαδικτύου | Web Object |
| Αξιολόγηση από Ομότιμους Κριτές | Peer Review |
| Απήχηση | Impact |
| Ασαφές Σύνολο | Fuzzy Set |
| Αυτοαναφορά | Self Citation |
| Βαθμολογία | Score |
| Βαθμός Εισερχόμενων Ακμών | In-Degree |
| Βαθμός Εξερχόμενων Ακμών | Out-Degree |
| Βιβλιομετρία | Bibliometrics |
| Βραχύχρονη Απήχηση | Short-term Impact |
| Δημοφιλία | Popularity |
| Δημοσιεύσεις, ή Εξαφανίζεσαι | Publish or Perish |
| Διάγραμμα Διασποράς | Scatter Plot |
| Διάγραμμα Ραντάρ | Radar Chart |
| Διαχείριση Δεδομένων και Γνώσης | Data and Knowledge Management |
| Δίκτυο Αναφορών | Citation Network |
| Δυναμομέθοδος | Power Method |
| Εικονικό Μηχάνημα | VM (Virtual Machine) |
| Εκθετικός Νόμος | Power Law |
| Εκκρεμής Κόμβος | Dangling Node |
| Ενημερωτικά Γραφικά | Infographics |
| Εξόρυξη Κειμένου | Text Mining |
| Επιστήμες Ζωής | Life Sciences |
| Επιστημομετρία | Scientometrics |
| Επιστημονικές Δημοσιεύσεις | Scientific Publications |
| Επιστημονικό Περιοδικό | Scientific Journal |
| Ετικέτα XML | XML Tag |
| Ισχυρά Συνδεδεμένος Γράφος | Strongly Connected Graph |
| Καθυστέρηση Αναφοράς | Citation Lag |
| Κατάταξη | Ranking |
| $k$-σημαντικότερα | Top-$k$ |
| Κεντρικότητα | Centrality |
| Κεντρικότητα Katz | Katz Centrality |
| Κλιμάκωση | Scalability |
| Κρίση Χρήστη | User Judgement |

| | |
|---|---|
| Λανθάνον Θέμα | Latent Topic |
| Μακρόχρονη Απήχηση | Long-term Impact |
| Μεγάλη Ουρά (Κατανομής) | Long Tail |
| Μέθοδοι Κατάταξης | Ranking Methods |
| Μεταδεδομένα | Metadata |
| Μέτρο Κεντρικότητας | Centrality Metric |
| Μηχανική Μάθηση | Machine Learning |
| Μηχανισμός Προσοχής | Attention Mechanism |
| Μικροϋπηρεσίες | Microservices |
| Μοντελοποίηση Θεμάτων | Topic Modelling |
| Παράγοντες Διάδοσης Δημοφιλίας | Popularity Propagation Factors |
| Παρακτρατημένα Δεδομένα | Held-out Data |
| Περιγραφική Αξιολόγηση | Descriptive Evaluation |
| Πιθανότητα Επιλογής | Landing Probability |
| Πιθανότητα Τυχαίου Άλματος | Random Jump Probability |
| Πίνακας Γειτνίασης | Adjacency Matrix |
| Πίνακας Μεταβάσεων | Transition Matrix |
| Προγραμματιστική Διεπαφή Εφαρμογών | Application Programming Interface (API) |
| Προτιμησιακή Προσκόλληση | Preferential Attachment |
| Σημασία ανεξαρτήτως Ερωτήματος | Query-independent Importance |
| Συνολική Επίδραση | Influence |
| Σύνολο Δεδομένων | Dataset |
| Σύστημα Προτάσεων | Recommender |
| Σχετικός με το Ερώτημα | Query-dependent |
| Τιμή Βάσης | Default |
| Τυχαία Περιήγηση | Random Walk |
| Τυχαίος Εντοπιστής Αντικειμένων | Random Object Finder |
| Τυχαίος Ερευνητής | Random Researcher |
| Υπόβαθρο Αληθείας | Ground Truth |
| Χάρτης Θερμότητας | Heatmap |
| Χρονικά Ενήμερος | Time Aware |

# Στοιχεία Επικοινωνίας

Εργαστήριο Συστημάτων Γνώσης και Δεδομένων
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Εθνικό Μετσόβιο Πολυτεχνείο
Ηρώων Πολυτεχνείου 9. Ζωγράφου
157 80 Αθήνα, Ελλάδα
Τηλέφωνο: (+30) 210 772 1402
Ηλεκτρονικό Ταχυδρομείο (e-mail): kanellos@dblab.ece.ntua.gr

Ινστιτούτο Πληροφοριακών Συστημάτων
Ερευνητικό Κέντρο «ΑΘΗΝΑ»
Αρτέμιδος 6 και Επιδαύρου, Μαρούσι
15 125 Αθήνα, Ελλάδα
Ηλεκτρονικό Ταχυδρομείο (e-mail): ilias.kanellos@athenarc.gr

# Σπουδές, επαγγελματικές άδειες και πιστοποιήσεις

- 2012–Σήμερα: Υποψήφιος διδάκτορας Εθνικού Μετσόβιο Πολυτεχνείου, σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών (επιβλέπων καθηγητής: Ιωάννης Βασιλείου)

- 2005–2012: Προπτυχιακός φοιτητής στο τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου (βαθμός διπλώματος: 7,94/10)

- Άδεια ασκήσεως του επαγγέλματος του Ηλεκτρολόγου Μηχανικού και Μηχανικού Υπολογιστών από το Τεχνικό Επιμελητήριο Ελλάδος (ΤΕΕ)

- Απολυτήριο λυκείου από το 8ο ενιαίο λύκειο Ηρακλείου Κρήτης

- Πιστοποιητικά γνώσης της Αγγλικής γλώσσας σε επίπεδο επάρκειας Proficiency από το University of Cambridge και το University of Michigan

- Πιστοποιητικό γνώσης της Γερμανικής γλώσσας σε ανώτατο επίπεδο (C2) από το Goethe Institut

# Ερευνητικά Ενδιαφέροντα

- Αλγόριθμοι κατάταξης επιστημονικών δημοσιεύσεων

- Αυτόματη εξαγωγή γνώσης από επιστημονικές δημοσιεύσεις με εξόρυξη κειμένου.

- Διαχείριση Επιστημονικών Δεδομένων

- Υπολογισμοί νέφους (συστήματα Hadoop-Spark)

# Επαγγελματική Εμπειρία

- Επιστημονικός συνεργάτης ΙΠΣΥ, Ε.Κ. «ΑΘΗΝΑ» την περίοδο 2013-2019.
- Πρακτική άσκηση στο Εργαστήριο Πληροφοριακών Συστημάτων του Τμήματος Πληροφορικής στο Ίδρυμα Τεχνολογίας και Έρευνας την περίοδο 01/07/2010 ως 30/9/2010.

# Εκπαιδευτική Εμπειρία

- Συμμετοχή στην επίβλεψη των ακόλουθων διπλωματικών εργασιών:
  - Υλοποίηση μηχανισμού κατάταξης δημοσιεύσεων για δημοσιεύσεις σχετικές με microRNA, Βασιλική Βλαχοκυριάκου, 2015.
  - Σύστημα για την αυτόματη εξαγωγή αλληλεπιδράσεων microRNA-γονιδίων από επιστημονικές δημοσιεύσεις στις βιοεπιστήμες, Ροδοθέα-Μυρσίνη Τσουπίδι, 2014.
  - Σύστημα για τη διαχείριση εξελισσόμενων γονιδιακών δεδομένων, Κωνσταντίνος Ζαγγανάς, 2014.

# Επιστημονικές δημοσιεύσεις

**Περιοδικά με Κρίση:**

- Kanellos I, Vergoulis T, Sacharidis D, Dalamagas T, Vassiliou Y. Impact-Based Ranking of Scientific Publications: A Survey and Experimental Evaluation. IEEE Transactions on Knowledge and Data Engineering. 2019 Sep 13.
- Karagkouni D, Paraskevopoulou MD, Chatzopoulos S, Vlachos IS, Tastsoglou S, Kanellos I, Papadimitriou D, Kavakiotis I, Maniou S, Skoufos G, Vergoulis T. DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA–gene interactions. Nucleic acids research. 2017 Nov 16;46(D1):D239-45.
- Georgakilas G, Vlachos IS, Zagganas K, Vergoulis T, Paraskevopoulou MD, Kanellos I, Tsanakas P, Dellis D, Fevgas A, Dalamagas T, Hatzigeorgiou AG. DIANA-miRGen v3. 0: accurate characterization of microRNA promoters and their regulators. Nucleic acids research. 2016 Jan 4;44(D1):D190-5.
- Paraskevopoulou MD, Vlachos IS, Karagkouni D, Georgakilas G, Kanellos I, Vergoulis T, Zagganas K, Tsanakas P, Floros E, Dalamagas T, Hatzigeorgiou AG. DIANA-LncBase v2: indexing microRNA targets on non-coding transcripts. Nucleic acids research. 2016 Jan 4;44(D1):D231-8.
- Vergoulis T, Kanellos I, Kostoulas N, Georgakilas G, Sellis T, Hatzigeorgiou A, Dalamagas T. mirPub: a database for searching microRNA publications. Bioinformatics. 2015 May 1;31(9):1502-4.
- Vlachos IS, Paraskevopoulou MD, Karagkouni D, Georgakilas G, Vergoulis T, Kanellos I, Anastasopoulos IL, Maniou S, Karathanou K, Kalfakakou D,

Fevgas A. DIANA-TarBase v7. 0: indexing more than half a million experimentally supported miRNA: mRNA interactions. Nucleic acids research. 2015 Jan 28;43(D1):D153-9.

### Συνέδρια με Κρίση:

- Vergoulis T, Chatzopoulos S, Kanellos I, Deligiannis P, Tryfonopoulos C, Dalamagas T. BIP! Finder: Facilitating Scientific Literature Search by Exploiting Impact-Based Ranking. InProceedings of the 28th ACM International Conference on Information and Knowledge Management 2019 Nov 3 (pp. 2937-2940). ACM.
- Vergoulis T, Kanellos I, Tzerefos A, Chatzopoulos S, Dalamagas T, Skiadopoulos S. A Study on the Readability of Scientific Publications. InInternational Conference on Theory and Practice of Digital Libraries 2019 Sep 9 (pp. 136-144). Springer, Cham.
- Chatzopoulos S, Deligiannis P, Vergoulis T, Kanellos I, Tryfonopoulos C, Dalamagas T. SciTo trends: visualising scientific topic trends. InInternational Conference on Theory and Practice of Digital Libraries 2019 Sep 9 (pp. 393-396). Springer, Cham.
- Kanellos I, Vlachokyriakou V, Vergoulis T, Georgakilas G, Vassiliou Y, Hatzigeorgiou AK, Dalamagas T. MirPub v2: Towards Ranking and Refining miRNA Publication Search Results. InInternational Conference on Theory and Practice of Digital Libraries 2015 Sep 14 (pp. 355-359). Springer International Publishing.
- Tsoupidi RM, Kanellos I, Vergoulis T, Vlachos IS, Hatzigeorgiou AG, Dalamagas T. TarMiner: automatic extraction of miRNA targets from literature. InProceedings of the 27th International Conference on Scientific and Statistical Database Management 2015 Jun 29 (p. 12). ACM.
- Kanellos I, Vergoulis T, Sacharidis D, Dalamagas T, Hatzigeorgiou A, Sartzetakis S, Sellis T. MR-microT: a MapReduce-based MicroRNA target prediction method. InProceedings of the 26th International Conference on Scientific and Statistical Database Management 2014 Jun 30 (p. 47). ACM.

## Συμμετοχή σε Έργα

- Smart Data Lake, H2020 – ICT – 12 – 2018-2020. (2019)
- KAME – Καινοτόμες Μέθοδοι για την Ανάλυση και Διαχείριση Μεγάλων Δεδομένων (2017)
- SlideWiki – Large-scale pilots for collaborative OpenCourseWare authoring, multiplatform delivery and Learning Analytics, H2020 – ICT – 2015. (2016)
- ΜΕΔΑ – Μεγάλα Δεδομένα, Προκλήσεις, Μέθοδοι και Αποδοτικές Τεχνικές Διαχείρισης Δεδομένων (2014-2015)