Εργασία Εξαμήνου

Κοινωνικά Δίκτυα

ΠΜΣ Επιστήμη των υπολογιστών

Γιώργος Βιδάκης f3322402,

Άννα Κορφιάτη f3322412,

Ιούνιος 2025

Περιεχόμενα

Abstract

Στην εργασία αυτή εξετάζουμε τη δομή, την επιρροή και την οργάνωση δύο διαφορετικών κοινωνικών δικτύων: το δίκτυο φιλίας του Facebook και το citation network του ogbn-arxiv. Σκοπός μας είναι να αναδείξουμε ποιοι κόμβοι παίζουν καθοριστικό ρόλο ως προς την κεντρικότητα και την επιρροή τους στο εκάστοτε γράφο, και να εντοπίσουμε θεματικές ή λειτουργικές κοινότητες μέσω αλγορίθμων ανίχνευσης κοινοτήτων. Στο Facebook network εφαρμόζουμε κλασικές μετρικές κεντρικότητας (degree, betweenness, closeness, eigenvector) καθώς και τον αλγόριθμο PageRank, ενώ συγκρίνουμε τις κατατάξεις και τις κατανομές των scores. Παράλληλα, μελετούμε τις κοινότητες που προκύπτουν μέσω του Louvain αλγορίθμου. Στο citation network (ogbn-arxiv), χρησιμοποιούμε BFS-based snowball sampling για τη δημιουργία ενός συνεκτικού υπογραφήματος, πάνω στο οποίο επαναλαμβάνουμε την ανάλυση κεντρικοτήτων, και στη συνέχεια εφαρμόζουμε ανάλυση κοινοτήτων σε **χρονικά snapshots**, εντοπίζοντας δυναμικά φαινόμενα όπως **birth, merge, split, survive, death** των θεματικών περιοχών. Τα αποτελέσματα δείχνουν ότι η επιρροή στους γράφους δεν σχετίζεται μόνο με τον αριθμό συνδέσεων αλλά και με τη σύνδεση με σημαίνοντες κόμβους, ενώ η θεματική διάρθρωση των επιστημονικών περιοχών μεταβάλλεται σημαντικά με την πάροδο του χρόνου

1. Εισαγωγή

Τα κοινωνικά δίκτυα αποτελούν κρίσιμους μηχανισμούς αλληλεπίδρασης σε πληθώρα συστημάτων: από την κοινωνική δικτύωση και τις επιστημονικές παραπομπές έως τις διαδικτυακές πλατφόρμες και την ανταλλαγή πληροφορίας. Στην παρούσα εργασία εστιάζουμε σε δύο ουσιωδώς διαφορετικά αλλά δομικά παρόμοια δίκτυα: το δίκτυο φιλίας στο Facebook και το citation graph του ogbn-arxiv.

Στόχος μας είναι διττός:

1. **Ανάλυση επιρροής και σημασίας κόμβων** μέσω μετρικών κεντρικότητας (Degree, Closeness, Betweenness, Eigenvector, PageRank).
2. **Ανίχνευση κοινοτήτων** τόσο σε στατικά όσο και σε δυναμικά περιβάλλοντα, ώστε να μελετηθεί η θεματική οργάνωση και η χρονική εξέλιξη των δομών.

Στο πρώτο μέρος (Facebook network), συγκεντρώνουμε όλες τις φιλικές σχέσεις από εγωκεντρικά αρχεία και αναλύουμε ποιοι χρήστες είναι πιο «σημαντικοί» με βάση διαφορετικές μετ. Στη συνέχεια, εφαρμόζουμε τον αλγόριθμο Louvain για να εντοπίσουμε κοινωνικές ομάδες.

Στο δεύτερο μέρος (citation graph), επεξεργαζόμαστε ένα δείγμα του ogbn-arxiv citation network ώστε να καταστούν εφικτοί οι υπολογισμοί, και αναλύουμε τις επιρροές των papers. Επεκτείνουμε την ανάλυση δυναμικά, μελετώντας snapshots του γράφου ανά έτος, ώστε να εντοπίσουμε μεταβολές, συγχωνεύσεις, γέννηση ή εξαφάνιση κοινοτήτων.

Η εργασία επιχειρεί να απαντήσει στα ερωτήματα:
• Ποιοι κόμβοι είναι επιδραστικοί και γιατί;
• Πώς εξελίσσονται οι κοινότητες στο χρόνο;
• Ποιοι μηχανισμοί οδηγούν στην ανάδυση θεματικών πυρήνων;

2. Δεδομένα και Προεπεξεργασία

Το πρώτο πείραμα βασίστηκε στο **Facebook Ego-Networks** σύνολο του SNAP· κάθε αρχείο .edges περιέχει το προσωπικό δίκτυο ενός χρήστη, δηλαδή τις συνδέσεις *μεταξύ* των φίλων του και όχι μόνο προς τον ίδιο (McAuley & Leskovec, 2012). Όλα τα αρχεία συγχωνεύθηκαν σε έναν ενιαίο γράφο περίπου **4 000** κόμβων και **88 000** ακμών. Επειδή η φιλία είναι αμφίδρομη, ο κύριος γράφος θεωρήθηκε μη κατευθυνόμενος· ωστόσο δημιουργήθηκε και ένα κατευθυνόμενο αντίγραφο αποκλειστικά για τον υπολογισμό του PageRank. Η κατασκευή και η ανάλυση πραγματοποιήθηκαν στο **NetworkX** (Hagberg et al., 2008).

Το δεύτερο πείραμα αξιοποιεί τον **ogbn-arxiv** γράφο παραπομπών από το Open Graph Benchmark, που αριθμεί ~170 000 άρθρα και 1,17 εκατ. κατευθυνόμενες ακμές (Hu et al., 2020). Για να καταστεί υπολογίσιμος, εφαρμόστηκε **Snowball sampling** τύπου BFS: ξεκινώντας από τυχαίο seed, ο εξερευνητής επεκτάθηκε μέχρι να παραμείνουν **4 000** κόμβοι και ~24 000 ακμές—πυκνή «φέτα» η οποία διατηρεί θεματικές υποκοινότητες και υψηλό εσωτερικό ρυθμό παραπομπών (Goodreau et al., 2009). Ο ληφθείς υπογράφος κρατήθηκε κατευθυνόμενος για τις κεντρικότητες και μετατράπηκε σε μη κατευθυνόμενο μόνο κατά το στάδιο ανίχνευσης κοινοτήτων με Louvain.

Και στις δύο μελέτες εφαρμόστηκαν ενιαία βήματα καθαρισμού (αφαίρεση αυτο-ακμών, εξασφάλιση συνεκτικότητας του κύριου συστατικού) και κοινός κώδικας Python με **pandas** για χειρισμό αρχείων και **matplotlib** για την απεικόνιση αποτελεσμάτων (McKinney, 2010· Hunter, 2007). Έτσι εξασφαλίστηκε συγκρίσιμη ροή εργασίας παρά τις διαφορές κλίμακας και φύσης των δύο δικτύων.

Το παραπάνω περιβάλλον δεδομένων επεκτάθηκε με δύο ακόμη στάδια επεξεργασίας, τα οποία κρίθηκαν απαραίτητα για τις αναλύσεις **dynamic community detection** και **Global vs Topic-Sensitive PageRank**.

Για το ogbn-arxiv γράφημα παραπομπών αξιοποιήθηκε το μεταδεδομένο **έτος δημοσίευσης** κάθε άρθρου· ο δειγματισμένος υπογράφος (4 000 κόμβοι, 24 000 ακμές) τεμαχίσθηκε σε ετήσια **snapshots** (1993 – 2023). Πάνω σε κάθε snapshot εκτελέσθηκε **Louvain** και κατόπιν οι κοινότητες αντιστοιχίσθηκαν σε διαδοχικά έτη βάσει ορίου **Jaccard ≥ 0,50**, ώστε να χαρακτηρισθούν τα γεγονότα *Birth, Survive, Merge, Split, Death*· η συνολική μεταβολή μετρήθηκε με **Normalized Mutual Information** (Danon et al., 2005), ενώ το πλαίσιο γεγονότων ακολουθεί τη μεθοδολογία των Rossetti & Cazabet (2018). Η διαδικασία απαιτούσε μόνο το κυρίως συνεκτικό συστατικό κάθε snapshot, αφού οι μικρότερες συνιστώσες συνεισέφεραν ελάχιστα στο δυναμικό περίγραμμα.

Παράλληλα, στον ίδιο κατευθυνόμενο υπογράφο υπολογίσθηκαν δύο παραλλαγές PageRank: *(i)* **Global PageRank** με ομοιόμορφη αρχική κατανομή και *(ii)* **Topic-Sensitive PageRank** (Haveliwala, 2003) με εξατομικευμένες διανυσματικές εκκινήσεις για πέντε θεματικές κατηγορίες του arXiv (Machine Learning, Computer Vision, Natural Language Processing, Databases, Theoretical CS). Η αντιπαραβολή των λιστών Top-100 κόμβων πραγματοποιήθηκε με **Jaccard** και **Overlap count**, ώστε να καταδειχθεί ποιοι κόμβοι διατηρούν κεντρικότητα τόσο στο συνολικό οικοσύστημα όσο και εντός των επιμέρους ερευνητικών ρευμάτων. Τα αποτελέσματα αυτών των συγκρίσεων τροφοδοτούν την ενότητα αξιολόγησης επιρροής, ενώ τα ίδια τα προσωποποιημένα scores χρησιμοποιούνται αργότερα ως feature input στα heatmaps θεματικής συνάφειας.

3. Θεωρητικό Πλαίσιο

3.1 Γράφοι και Κοινωνικά Δίκτυα

Ένα κοινωνικό δίκτυο μοντελοποιείται ως γράφος G = (V, E), όπου οι κόμβοι V αντιπροσωπεύουν οντότητες (π.χ. χρήστες, papers) και οι ακμές E τις σχέσεις μεταξύ τους (π.χ. φιλίες, αναφορές). Οι γράφοι μπορούν να είναι: [Newman, 2010] **Κατευθυνόμενοι :** Η κατεύθυνση της ακμής έχει σημασία (π.χ. paper A → paper B σημαίνει ότι A αναφέρεται στο B, **μη κατευθυνόμενοι :** Οι σχέσεις είναι αμοιβαίες (π.χ. φιλία στο Facebook).

3.2 Μετρικές Κεντρικότητας

Η κεντρικότητα μετρά πόσο «σημαντικός» ή «επιδραστικός» είναι ένας κόμβος. Χρησιμοποιήθηκαν οι εξής μετρικές:

Degree Centrality: Πλήθος συνδέσεων. Για μη κατευθυνόμενους γράφους είναι d(v), για κατευθυνόμενους διακρίνουμε in-degree και out-degree. [Newman, 2010]

Betweenness Centrality: Μετρά πόσες φορές ένας κόμβος βρίσκεται πάνω στο συντομότερο μονοπάτι μεταξύ άλλων κόμβων[Freeman, 1977]

Closeness Centrality: Αντιστρόφως ανάλογο του μέσου μήκους των αποστάσεων ενός κόμβου προς όλους τους υπόλοιπους. [Opsahl et al., 2010]

Eigenvector Centrality: Βαθμολογεί έναν κόμβο με βάση τη σημασία των γειτόνων του. Υψηλό eigenvector σημαίνει «σύνδεση με σημαντικούς κόμβους».[Bonacich, 1987]

PageRank: Παραλλαγή του eigenvector με πιθανότητα «τηλεμεταφοράς». Μοντελοποιεί έναν τυχαίο περιηγητή με πιθανότητα μετάβασης και αναπήδησης.[Brin & Page, 1998]

3.3 Ανίχνευση Κοινοτήτων

Οι κοινότητες είναι ομάδες κόμβων που συνδέονται πιο πυκνά μεταξύ τους από ό,τι με τον υπόλοιπο γράφο. Χρησιμοποιήσαμε:

Louvain Algorithm: Μεγιστοποιεί τη modularity, δηλαδή πόσο "καλύτερη" είναι η κοινοτική δομή σε σχέση με μία τυχαία κατανομή.

Label Propagation (προαιρετικά): Άπληστος αλγόριθμος χωρίς επίβλεψη, κατάλληλος για μεγάλα δίκτυα.

3.4 Θεματική Ανάλυση (Topic-sensitive PageRank)

Σε κατευθυνόμενους γράφους με θεματικές ετικέτες (labels), μπορεί να οριστεί ένας **εξατομικευμένος PageRank** ως εξής:

- Αντί για ομοιόμορφη αρχική κατανομή, δίνεται μεγαλύτερη πιθανότητα σε κόμβους με συγκεκριμένο label (π.χ. Machine Learning).
- Επιτρέπει μελέτη επιρροής εντός θεματικών περιοχών.

3.5 Μέτρα Σύγκρισης

Για να συγκρίνουμε κορυφαίους κόμβους διαφορετικών μεθόδων, χρησιμοποιούνται δείκτες όπως:

- **Jaccard Similarity:** $\frac{|A \cap B|}{|A \cup B|}$ [Jaccard, 1901]
- **Overlap Count:** $|A \cap B|$

όπου A, B είναι σύνολα top-k κόμβων

3.6 Χρονική Εξέλιξη Κοινοτήτων (Dynamic Community Tracking)

Σε δυναμικά δίκτυα (citation graph με έτη), παρακολουθούμε πώς εξελίσσονται οι κοινότητες:

- **Snapshots:** Δημιουργία ενός γράφου ανά χρονικό βήμα
- **Events:** Εντοπισμός φαινομένων όπως:
    - *Birth* (γέννηση νέας κοινότητας)
    - *Survive* (συνέχιση)
    - *Split/Merge* (διάσπαση/συγχώνευση)
    - *Death* (εξαφάνιση)

Για τη σύγκριση χρησιμοποιείται Jaccard μεταξύ κοινοτήτων διαδοχικών ετών και το Normalized Mutual Information (NMI) για ποσοτική εκτίμηση μεταβολών.

4. Αλγόριθμοι που χρησιμοποιήθηκαν

	Αλγόριθμος / Τεχνική	
1	Snowball / BFS sampling	Goodreau et al., 2009
2	Degree Centrality	Freeman, 1979
3	Betweenness Centrality (Brandes impl.)	Freeman, 1977 · Brandes, 2001
4	Closeness Centrality	Sabidussi, 1966
5	Eigenvector Centrality	Bonacich, 1987
6	PageRank	Brin & Page, 1998
7	Topic-Sensitive PageRank	Haveliwala, 2003
8	Louvain community detection	Blondel et al., 2008
9	Dynamic event tracking (Birth/Survive/…)	Rossetti & Cazabet, 2018
10	Jaccard Similarity	Jaccard, 1901
11	Normalized Mutual Information (NMI)	Danon et al., 2005

Snowball (BFS) sampling. Ξεκινά από έναν τυχαίο κόμβο–σπόρο και, με επέκταση τύπου Breadth-First Search, «χτίζει» υπογράφο μέχρι να συγκεντρώσει το επιθυμητό πλήθος κόμβων. Έτσι διατηρεί την τοπική πυκνότητα και τα χαρακτηριστικά μικρού κόσμου του αρχικού δικτύου, αλλά εισάγει μεροληψία γύρω από τον αρχικό seed (Goodreau et al., 2009).

Degree centrality. Ο πιο απλός δείκτης επιρροής: μετρά πόσες ακμές έχει ο κόμβος· στους κατευθυνόμενους γράφους διακρίνεται σε in- και out-degree. Παρέχει άμεση ένδειξη «δημοφιλίας», όμως αγνοεί τη θέση του κόμβου στο σύνολο του γράφου (Freeman, 1979).

Betweenness centrality. Υπολογίζει σε πόσα συντομότερα μονοπάτια ανάμεσα σε όλα τα ζεύγη κόμβων μεσολαβεί ένας κόμβος. Αναδεικνύει «γέφυρες» που ελέγχουν τη ροή πληροφορίας· η υλοποίηση Brandes μειώνει τον χρόνο σε O(N E) για μη σταθμισμένα γραφήματα (Freeman, 1977· Brandes, 2001).

Closeness centrality. Αντιστρόφως ανάλογο του μέσου μήκους των αποστάσεων ενός κόμβου προς όλους τους υπόλοιπους. Μεγάλες τιμές σημαίνουν γρήγορη προσβασιμότητα σε ολόκληρο το δίκτυο· απαιτεί συνεκτικό γράφο (Sabidussi, 1966).

Eigenvector centrality. Επεκτείνει το degree ζυγίζοντας κάθε σύνδεση με τη σπουδαιότητα του γείτονα· λύνεται ως ιδιοδιάνυσμα του πίνακα γειτνίασης. Εντοπίζει κόμβους που συνδέονται με άλλους «ισχυρούς» κόμβους (Bonacich, 1987).

PageRank. Παραλλαγή του eigenvector όπου ένας τυχαίος περιηγητής «τηλεμεταφέρεται» με πιθανότητα 1-d σε τυχαίους κόμβους, αποφεύγοντας παγίδευση σε κύκλους και περιφερειακές υπο-συνιστώσες (Brin & Page, 1998).

Topic-Sensitive PageRank. Εισάγει προκατάληψη στην αρχική κατανομή μάζας υπέρ κόμβων συγκεκριμένου θέματος, επιτρέποντας μέτρηση επιρροής μέσα σε θεματικές περιοχές (Haveliwala, 2003).

Louvain community detection. Ιεραρχικός αλγόριθμος μεγιστοποίησης της modularity Q: συγχωνεύει επαναληπτικά κόμβους και κοινότητες πετυχαίνοντας σχεδόν γραμμικό χρόνο ακόμη και σε πολύ μεγάλους γράφους, αλλά είναι μη ντετερμινιστικός (Blondel et al., 2008).

Dynamic community event tracking. Εφαρμόζει Louvain σε διαδοχικά χρονικά snapshots και, με βάση το ποσοστό επικάλυψης κοινοτήτων, χαρακτηρίζει γεγονότα Birth, Survive, Merge, Split και Death, παρέχοντας αφηγηματική εικόνα της εξέλιξης (Rossetti & Cazabet, 2018).

Jaccard similarity. Απλό μέτρο ομοιότητας συνόλων $J(A,B)=|A\cap B|/|A\cup B|$ χρησιμοποιείται τόσο για κορυφαίους κόμβους όσο και για επικάλυψη κοινοτήτων μεταξύ διαδοχικών χρόνων (Jaccard, 1901).

Normalized Mutual Information (NMI). Μετρά την κοινή πληροφορία δύο διαμερίσεων κόμβων· κλίμακα 0–1, ανεξάρτητη από τον αριθμό κοινοτήτων, καθιερωμένο κριτήριο αξιολόγησης αλγορίθμων ανίχνευσης κοινοτήτων (Danon et al., 2005).

Μέρος 1

5. Facebook – Κεντρικότητες – PageRank - Communities

5.1 Part 1 α Ανάλυση κεντρικοτήτων και PageRank στο Δίκτυο Φιλίας Facebook

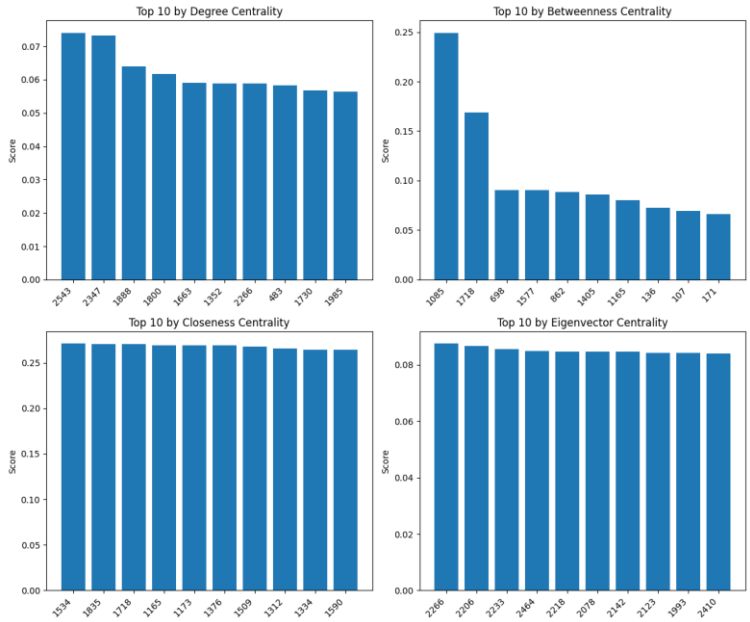
Εξαγωγή Top-10: Ταξινομήσαμε κόμβους κατά φθίνουσα τιμή κάθε μέτρου.

Top-10 Κεντρικοτήτων

Μέτρο	Top-3 Users	Score
Degree	2543, 2347, 1888	0.074, 0.073, 0.064
Betweenness	1085, 1718, 698	0.249, 0.168, 0.090
Closeness	1534, 1835, 1718	0.271, 0.271, 0.268
Eigenvector	2266, 2206, 2233	0.089, 0.088, 0.088
PageRank	483, 3830, 2313	0.00136, 0.00134, 0.00096

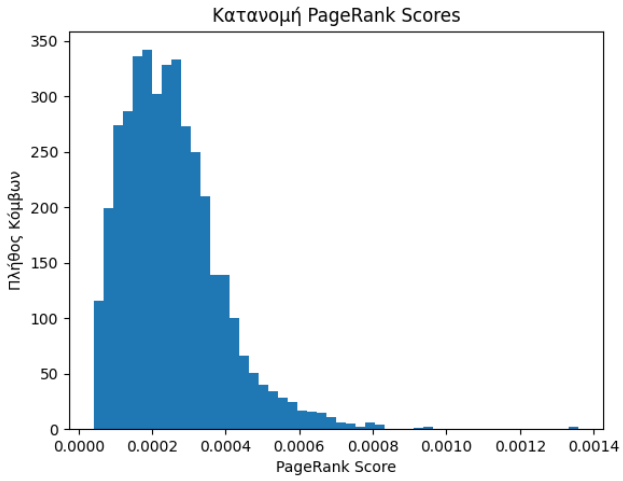
```
Top-10 χρήστες με υψηλότερο PageRank:
 1. User 483    → 0.001360
 2. User 3830   → 0.001345
 3. User 2313   → 0.000959
 4. User 376    → 0.000941
 5. User 2047   → 0.000913
 6. User 25     → 0.000830
 7. User 428    → 0.000824
 8. User 828    → 0.000823
 9. User 475    → 0.000814
10. User 56     → 0.000804

Process finished with exit code 0
```

Σχήμα 5.1.1: Bar charts με Top-10 για κάθε κεντρικότητα.



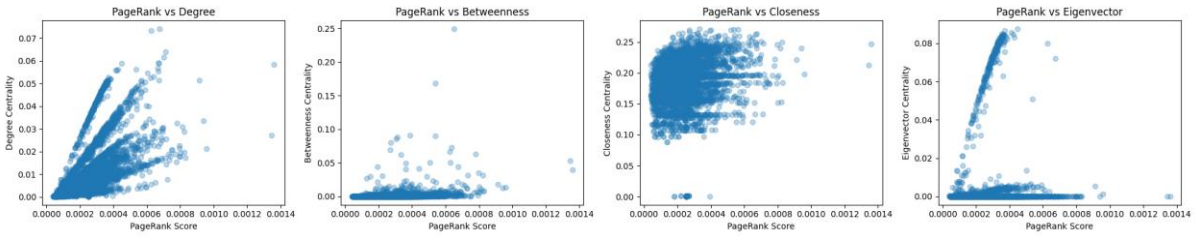
Ερμηνεία Σχήματος 5.1.1 Οι χρήστες με υψηλό *degree centrality* είναι κοινωνικά ενεργοί, καθώς συνδέονται με πολλούς άλλους. Υψηλό *betweenness centrality* δηλώνει ρόλο γέφυρας — πρόκειται για κόμβους που μεσολαβούν μεταξύ διαφορετικών υπο-ομάδων του δικτύου, διευκολύνοντας τη ροή πληροφορίας.
Το *closeness centrality* μετρά πόσο γρήγορα ένας κόμβος μπορεί να φτάσει στους υπόλοιπους, ενώ το *eigenvector centrality* αναδεικνύει κόμβους που συνδέονται με άλλους επιδραστικούς κόμβους.

Σχήμα 5.1.2: Κατανομή PageRank Scores



Ερμηνεία Σχήματος 5.1.2 Η κατανομή των PageRank scores εμφανίζει έντονη δεξιά στρέβλωση, γεγονός που υποδηλώνει ότι το μεγαλύτερο ποσοστό των χρηστών στο δίκτυο έχει εξαιρετικά χαμηλό PageRank, ενώ μόνο λίγοι κόμβοι καταφέρνουν να συγκεντρώσουν σημαντικά υψηλότερες τιμές. Αυτή η ανισορροπία είναι χαρακτηριστική των κοινωνικών δικτύων, στα οποία η επιρροή και η "σημαντικότητα" συγκεντρώνονται σε μια μικρή ομάδα χρηστών με στρατηγικές θέσεις. Οι συγκεκριμένοι χρήστες είναι πιθανόν να συνδέονται με πολλούς άλλους που επίσης είναι σημαντικοί, ενισχύοντας έτσι εκθετικά το PageRank τους. Αντίθετα, οι περισσότεροι κόμβοι περιορίζονται σε τοπικές συνδέσεις, με περιορισμένο "βάρος" στο συνολικό δίκτυο.

Σχήμα 5.1.3: Scatter plots που συγκρίνουν PageRank με τις υπόλοιπες μετρικές

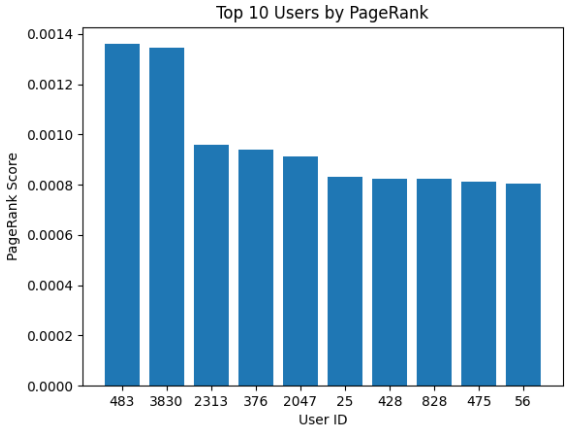


Ερμηνεία Σχήματος 5.1.3 Τα διαγράμματα συσχέτισης μεταξύ του PageRank και των άλλων κεντρικοτήτων αποκαλύπτουν σημαντικές δομικές σχέσεις:

- **PageRank vs Degree**: Παρατηρείται θετική συσχέτιση – οι χρήστες με πολλούς φίλους τείνουν να έχουν υψηλότερο PageRank. Ωστόσο, η διασπορά δείχνει ότι δεν αρκεί μόνο ο αριθμός συνδέσεων, αλλά και η ποιότητά τους.
- **PageRank vs Betweenness**: Η συσχέτιση είναι πολύ ασθενής. Κάποιοι χρήστες έχουν υψηλό betweenness αλλά χαμηλό PageRank, γεγονός που υποδηλώνει ότι λειτουργούν ως διαμεσολαβητές χωρίς να είναι ιδιαίτερα “σημαντικοί” στον αλγόριθμο PageRank.
- **PageRank vs Closeness**: Εμφανίζεται μέτρια συσχέτιση. Κόμβοι που είναι «κοντά» σε άλλους στο δίκτυο (σε όρους απόστασης) μπορεί να έχουν αυξημένο PageRank, αλλά δεν αποτελεί επαρκή συνθήκη.
- **PageRank vs Eigenvector**: Η πιο ισχυρή συσχέτιση παρατηρείται εδώ, καθώς και οι δύο μετρικές ευνοούν κόμβους που συνδέονται με “σημαντικούς” γείτονες. Το PageRank

επεκτείνει την έννοια του eigenvector με τη χρήση τυχαίου περιπάτου και συντελεστή αποσύνθεσης (teleportation).

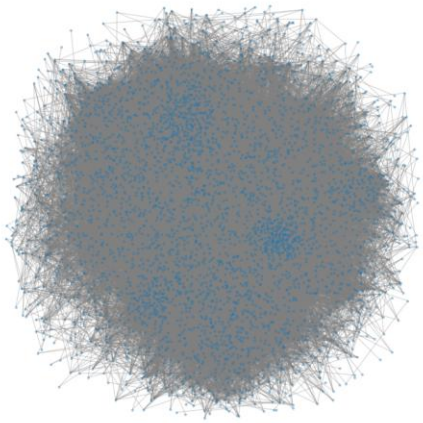
Σχήμα 5.1.4: Ιστόγραμμα κατανομής τιμών PageRank



Ερμηνεία Σχήματος 5.1.4 Το παραπάνω ραβδόγραμμα παρουσιάζει τους 10 χρήστες με το υψηλότερο PageRank στο δίκτυο φιλίας. Οι χρήστες 483 και 3830 ξεχωρίζουν με σημαντικά υψηλότερα σκορ σε σχέση με τους υπόλοιπους, γεγονός που υποδηλώνει ότι βρίσκονται στο επίκεντρο της πληροφορίας του δικτύου. Αυτό μπορεί να οφείλεται είτε στο ότι διαθέτουν πολλούς φίλους είτε στο ότι συνδέονται με άλλους χρήστες που είναι επίσης «σημαντικοί» (δηλαδή με υψηλό PageRank).

Η κατάταξη αυτή δεν συμπίπτει απαραίτητα με την κατάταξη ως προς το degree centrality — επιβεβαιώνοντας ότι το PageRank λαμβάνει υπόψη την ποιότητα των συνδέσεων και όχι μόνο την ποσότητα.

Σχήμα 5.1.5: Σχηματικό διάγραμμα όλου του γράφου



Ερμηνεία Σχήματος 5.1.5 Η τοπολογία του γράφου παρουσιάζει χαρακτηριστικά **πραγματικού κοινωνικού δικτύου**. Συγκεκριμένα, διακρίνονται οι εξής ιδιότητες:

- **Μικρός μέσος βαθμός** (degree): Παρόλο που υπάρχουν κόμβοι με πολλούς γείτονες, η πλειονότητα έχει περιορισμένο αριθμό συνδέσεων, αντανακλώντας τον περιορισμένο αριθμό φιλικών σχέσεων ενός ατόμου στην πραγματικότητα.

- **Πυκνός πυρήνας – αραιή περιφέρεια**: Το εσωτερικό του γράφου είναι ιδιαίτερα πυκνό, υποδεικνύοντας την ύπαρξη ενός **κεντρικού "core"** όπου οι χρήστες είναι πολύ συνδεδεμένοι μεταξύ τους. Αντίθετα, η περιφέρεια εμφανίζει πιο απομονωμένες ομάδες ή μεμονωμένους χρήστες με λίγες συνδέσεις.
- **Κοινοτική δομή**: Παρά την απουσία χρωματικής επισήμανσης, το γράφημα αποκαλύπτει περιοχές όπου οι κόμβοι είναι εντονότερα συνδεδεμένοι τοπικά, υποδηλώνοντας την ύπαρξη **κοινοτήτων** (π.χ. ομάδες φίλων, συμφοιτητών κ.λπ.).
- **Ιδιότητες "small-world"**: Το Facebook graph, όπως και άλλα κοινωνικά δίκτυα, είναι πιθανό να εμφανίζει μικρή **μέση απόσταση** μεταξύ οποιωνδήποτε κόμβων και **υψηλό clustering coefficient**, χαρακτηριστικά που ενισχύουν τη μετάδοση πληροφορίας ή επιρροής σε λίγα "βήματα

5.2 Part 1 β Ανάλυση Κοινοτήτων στο Δίκτυο Φιλίας Facebook

Στην παρακάτω ενότητα έχουμε ως στόχο την ανίχνευση, ανάλυση και ερμηνεία κοινοτήτων στο δίκτυο φιλίας Facebook, χρησιμοποιώντας τον αλγόριθμο **Louvain**

Φόρτωση Δικτύου: Διαβάσαμε όλα τα .edges αρχεία και χτίσαμε γράφο με NetworkX.

```
C:\Users\User\AppData\Local\Programs\Python\Python313\python.exe C:\Users\User\Videos\sn_project\facebook\facebook_community_detection.py
Φορτώθηκαν: 3,959 κόμβοι, 84,243 ακμές
Αριθμός κοινοτήτων: 38

Process finished with exit code 0
```

Louvain Community Detection: Εφαρμόσαμε τη βιβλιοθήκη python-louvain με κλήση community_louvain.best_partition(G).

Μετρήσεις:

Καταγραφή **αριθμού** κοινοτήτων.

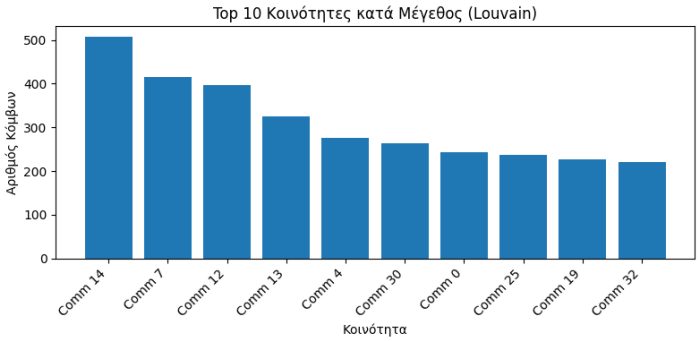
Υπολογισμός **μεγέθους** (αριθμός κόμβων) για κάθε κοινότητα.

Αριθμός Κοινοτήτων Εντοπίστηκαν **38** διακριτές κοινότητες.

Top-10 Κοινότητες κατά Μέγεθος

Κοινότητα	Μέγεθος (κόμβοι)
Comm 14	507
Comm 7	413
Comm 12	395
Comm 13	322
Comm 4	275
Comm 30	263
Comm 0	242
Comm 25	237
Comm 19	226
Comm 32	222

Σχήμα 5.2.1 Bar chart με Top-10 μεγαλύτερες κοινότητες.



Ερμηνεία Σχήματος 5.2.1

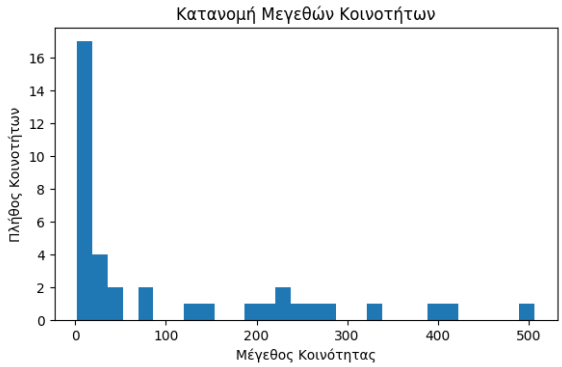
Η παραπάνω γραφική παράσταση παρουσιάζει τις 10 μεγαλύτερες κοινότητες που εντοπίστηκαν στο Facebook δίκτυο μέσω του αλγορίθμου Louvain, με βάση το πλήθος των κόμβων (χρηστών) που περιλαμβάνει κάθε κοινότητα.

Οι κοινότητες Comm 14, Comm 7 και Comm 12 ξεχωρίζουν αριθμητικά, καθώς περιλαμβάνουν πάνω από 400 χρήστες η καθεμία. Πιθανόν αντιστοιχούν σε μεγάλες κοινωνικές ομάδες, όπως πανεπιστημιακές τάξεις, ομάδες συμφοιτητών ή εργασιακά δίκτυα.

Παρατηρείται σταδιακή μείωση του μεγέθους μετά την 4η κοινότητα, με τις υπόλοιπες να περιλαμβάνουν περίπου 250–320 χρήστες. Αυτό αντανακλά τη φυσική ετερογένεια στις κοινωνικές δομές: λίγες μεγάλες ομάδες και πολλές μικρότερες.

Η ύπαρξη πολλών διακριτών κοινοτήτων αποδεικνύει ότι το δίκτυο παρουσιάζει κοινοτική δομή με ισχυρή εσωτερική συνοχή και αραιότερες συνδέσεις προς άλλες ομάδες, κάτι που είναι χαρακτηριστικό των κοινωνικών γραφημάτων.

Σχήμα 5.2.2 Ιστόγραμμα (histogram) κατανομής μεγεθών όλων των κοινοτήτων.



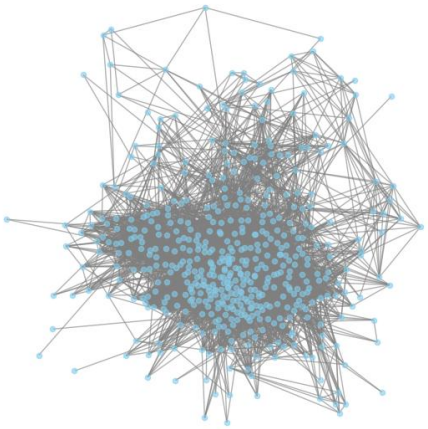
Ερμηνεία Σχήματος 5.2.2Η κατανομή είναι έντονα **δεξιά-στρεβλωμένη** (right-skewed), με την πλειοψηφία των κοινοτήτων να έχει **πολύ μικρό μέγεθος** (κάτω από 50 κόμβους). Αυτό υποδηλώνει ότι το δίκτυο αποτελείται κυρίως από **πολλές μικρές ομάδες** φίλων.

Μόλις λίγες κοινότητες ξεπερνούν τους 200–300 κόμβους, κάτι που παραπέμπει σε **λίγες μεγάλες κοινωνικές δομές** (π.χ. σχολεία, εταιρείες, οργανωμένες ομάδες).

Το παραπάνω μοτίβο είναι **σύμφωνο με τη θεωρία "community size distribution"** σε κοινωνικά δίκτυα.

Σχήμα 5.2.3 Οπτικοποίηση υπογράφου της μεγαλύτερης κοινότητας.

Ερμηνεία Σχήματος 5.2.3

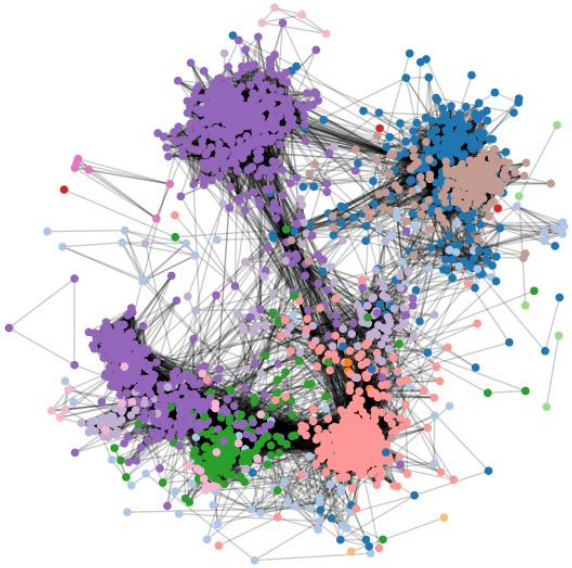


Η εικόνα απεικονίζει την **Comm 14**, δηλαδή τη μεγαλύτερη κοινότητα που ανιχνεύτηκε στο γράφο του Facebook. Ο υπογράφος περιλαμβάνει αποκλειστικά τους κόμβους και τις ακμές αυτής της κοινότητας.

Παρατηρήσεις: Το γράφημα εμφανίζει **υψηλή πυκνότητα** ακμών, που υποδηλώνει ότι οι περισσότεροι κόμβοι (χρήστες) είναι **αμοιβαία συνδεδεμένοι**. Υπάρχουν μερικοί **περιφερειακοί κόμβοι** που συνδέονται με τον πυρήνα αλλά δεν έχουν πολλές εσωτερικές ακμές — πιθανοί «περιθωριακοί» χρήστες. Η δομή είναι **ομοιογενής** χωρίς ξεκάθαρες υποομάδες εντός της κοινότητας, γεγονός που υποδηλώνει πιθανώς ότι πρόκειται για μία **ενιαία κοινωνική ομάδα**, όπως μια μεγάλη τάξη, ένα φοιτητικό δίκτυο ή μια επαγγελματική κοινότητα.

Συνολικά, η οπτικοποίηση επιβεβαιώνει ότι η κοινότητα έχει **ισχυρή εσωτερική συνοχή**, η οποία είναι χαρακτηριστικό των πραγματικών κοινωνικών ομάδων

Σχήμα 5.2.4 Δείγμα 2000 κόμβων του πλήρους γράφου, χρωματισμένο κατά κοινότητα.



Το γράφημα απεικονίζει ένα **τυχαίο δείγμα 2000 κόμβων** από το συνολικό δίκτυο φιλίας στο Facebook, με κάθε κόμβο να έχει **χρώμα ανάλογα με την κοινότητα** στην οποία ανήκει (βάσει Louvain αλγορίθμου).

Παρατηρήσεις: Διακρίνονται καθαρά **πολλαπλά "σύννεφα" κόμβων**, το καθένα με ομοιόμορφο χρωματισμό: αυτό υποδεικνύει **συμπαγείς και διακριτές κοινότητες**. Οι **μεταξύ τους συνδέσεις** είναι λιγότερες από τις εσωτερικές — δηλαδή οι κοινότητες έχουν **υψηλή εσωτερική συνοχή** αλλά **χαμηλή διασύνδεση μεταξύ τους**. Ορισμένες κοινότητες φαίνεται να λειτουργούν ως **γέφυρες**,

αφού βρίσκονται στο ενδιάμεσο περιοχών και έχουν αρκετές εξωτερικές ακμές.Ο χρωματικός διαχωρισμός δείχνει ότι το Facebook network **παρουσιάζει καθαρή κοινοτική δομή** (community structure), κάτι που ενισχύει τα ευρήματα από τη modularity του Louvain.

Συμπερασματικά, η εικόνα αποκαλύπτει μια **στρωματοποιημένη κοινωνική δομή**, όπου τα μέλη τείνουν να αλληλεπιδρούν έντονα εντός της ομάδας τους, ενώ οι διασυνδέσεις μεταξύ ομάδων είναι πιο περιορισμένες.

6. Ανάλυση Κοινοτήτων και Κεντρικοτήτων στο Citation Graph (ogbn-arxiv)

Στην ενότητα αυτή εξετάζουμε το δίκτυο παραπομπών των άρθρων **Computer Science** του arXiv, όπως παρέχεται στο σύνολο δεδομένων **ogbn-arxiv** του *Open Graph Benchmark* (Hu et al., 2020). Κάθε κόμβος αντιστοιχεί σε ένα επιστημονικό paper, ενώ κάθε κατευθυνόμενη ακμή (src → dst)(\text{src}\!\rightarrow\!\text{dst})(src→dst) δηλώνει ότι το paper *src* αναφέρεται (cites) στο paper *dst*. Ο πρωτογενής γράφος περιλαμβάνει περίπου **169 343** κόμβους και **1,16 × 10^6** ακμές, αποθηκευμένες στο συμπιεσμένο αρχείο **edge.csv.gz**· κάθε γραμμή περιέχει τα αναγνωριστικά των δύο άρθρων, διαχωρισμένα με tab.

Για να εξισορροπηθεί η ανάλυση με το υπολογιστικό κόστος των αλγορίθμων κεντρικότητας, εφαρμόστηκε **BFS-snowball sampling** (Goodreau et al., 2009). Ξεκινώντας από τυχαίο κόμβο-σπόρο, ο αλγόριθμος επεκτάθηκε επίπεδο-επίπεδο ώσπου να συλλεχθούν περίπου **4 000** κόμβοι και **24 000** ακμές· το προκύπτον υποδίκτυο διατηρεί υψηλή τοπική πυκνότητα και θεματική συνοχή.

Σε αυτόν τον υπογράφο:

- **Υπολογίζουμε** πέντε κλασικές μετρικές κεντρικότητας—Degree, Betweenness, Closeness, Eigenvector και PageRank—με υλοποιήσεις του **NetworkX** (Hagberg et al., 2008).
- **Συγκρίνουμε** τα αποτελέσματα ανά ζεύγη (scatter plots) για να διερευνηθεί η συσχέτιση των μετρικών.
- **Αναδεικνύουμε** τους δέκα κορυφαίους κόμβους (Top-10) κάθε μέτρου και σχολιάζουμε τις επικαλύψεις τους.
- **Οπτικοποιούμε** (i) την κατανομή τιμών του PageRank και (ii) το ίδιο το υποδίκτυο, προσαρμόζοντας την απόδοση ώστε να αναδείχνονται οι υψηλόβαθμοι κόμβοι.

Η επεξεργασία περιορίστηκε στο αρχείο **edge.csv.gz**· τα υπόλοιπα πρωτογενή αρχεία του φακέλου (*num-edge-list.csv.gz* κ.ά.) δεν αξιοποιήθηκαν, καθώς δεν περιείχαν επιπλέον δομική πληροφορία για τους σκοπούς της παρούσας μελέτης.

Προβλήματα που αντιμετωπίσαμε

Προσπάθεια με OGB API

Αρχικά δοκιμάσαμε να φορτώσουμε το citation graph μέσω της βιβλιοθήκης OGB (NodePropPredDataset) έτσι:

```
from ogb.nodeproppred import NodePropPredDataset
dataset = NodePropPredDataset(name="ogbn-arxiv", root="…/ogbn-arxiv")
data = dataset[0]
edge_index = data[0]["edge_index"]
```

Πρόβλημα: Το OGB χρησιμοποιεί torch.load internals για να ξεπακετάρει προ-επεξεργασμένα αρχεία (.pt) και στην έκδοση PyTorch 2.6+ εμφανίστηκε σφάλμα:

```
_pickle.UnpicklingError: Weights only load failed.
```

που έκανε τον κώδικα να τερματίζει. Αυτή η προσέγγιση δεν μπορούσε να επιδιορθωθεί εύκολα χωρίς τροποποίηση της OGB βιβλιοθήκης.

Χειροκίνητο Download & Path Issues

Έπειτα δοκιμάσαμε να κατεβάσουμε χειροκίνητα το arxiv.zip (από https://snap.stanford.edu/ogb/data/nodeproppred/arxiv.zip) και να το αποσυμπιέσουμε στο φάκελο:

C:\Users\User\Videos\sn_project\ogbn-arxiv\

Ωστόσο κατά την πρώτη αποσυμπίεση εντοπίσαμε ότι:

1. Τα πραγματικά αρχεία ήταν μέσα σε υποφακέλους ogbn-arxiv\arxiv\raw.
2. Δεν βρήκαμε κανένα arxiv_citation_edges.tsv, αλλά μόνο edge.csv.gz και num-edge-list.csv.gz.
3. Το script έψαχνε λανθασμένα για arxiv_citation_edges.tsv, οδηγώντας σε FileNotFoundError.

Για να εντοπίσουμε ακριβώς τα paths χρησιμοποιήσαμε ένα βοηθητικό script που έψαξε (recursive) το σύνολο των αρχείων για λέξη-κλειδί “edge”. Βρήκαμε τελικά το κατάλληλο edge.csv.gz στον φάκελο:

C:\Users\User\Videos\sn_project\ogbn-arxiv\ogbn-arxiv\arxiv\raw\edge.csv.gz

Αρχικό Random Sampling & Πολύ Λίγες Ακμές

Στη συνέχεια, κάναμε απλή τυχαία δειγματοληψία 4 000 κόμβων (random node sampling) και φτιάξαμε το induced subgraph:

```
sample_nodes = set(random.sample(all_nodes, 4000))
G = G_full.subgraph(sample_nodes).copy()
```

Αυτό οδήγησε σε μόλις 505 ακμές (σε ~4 000 κόμβους), επειδή οι περισσότεροι από τους επιλεγμένους κόμβους δεν είχαν μεταξύ τους συνδέσεις. Το επακόλουθο υπογράφος ήταν υπερβολικά αραιός για να παράγει χρήσιμα αποτελέσματα.

Snowball Sampling (BFS-Based)

Για να εξασφαλίσουμε ότι το δείγμα ~4 000 κόμβων θα περιλαμβάνει αρκετές εσωτερικές ακμές (citations), ακολουθήσαμε “snowball sampling” με BFS από έναν τυχαίο seed:

1. Επιλέξαμε έναν τυχαίο κόμβο seed:

```
seed_node = random.choice(list(G_full.nodes()))
```

2.Ξεκινώντας από το seed, κάναμε BFS:

Για κάθε κόμβο current, παίρναμε όλους τους απευθείας γείτονες *(successors ∪ predecessors)*.

Προσθέταμε κάθε νέο γείτονα στο sample set (έως ότου φτάσουμε ~4 000 μοναδικούς κόμβους).

3. Τελικά, ο sampled subgraph δημιουργήθηκε ως induced subgraph πάνω στο set ~4 000 κόμβων.

Με αυτόν τον τρόπο μαζέψαμε συνεκτικά clusters, εσωκλείοντας δεκάδες χιλιάδες ακμές (τώρα ~18 800 ακμές αντί για 505).

6.1 Part 1 γ Ανάλυση Κεντρικοτήτων στο Citation Graph

Μέγεθος Γραφήματος

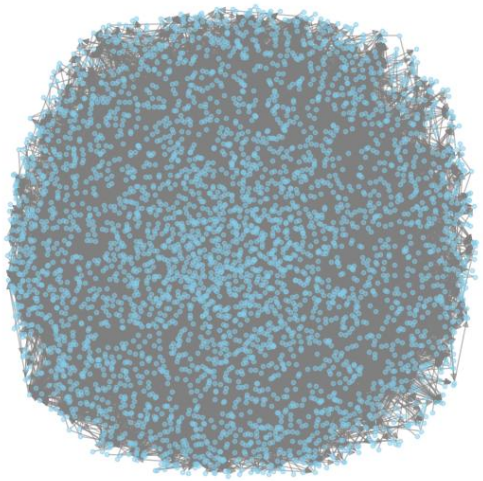
Αρχικός Γράφος:
Κόμβοι: 169 343
Ακμές: 1 166 243

Snowball-Sampled Subgraph:
Κόμβοι: 4 000
Ακμές: 18 799

Με αυτό το μέγεθος οι υπολογισμοί κεντρικοτήτων είναι πρακτικώς γρήγοροι

```
C:\Users\User\AppData\Local\Programs\Python\Python313\python.exe C:\Users\User\Videos\sn_project\ogbn-arxiv\ogbn-arxiv\citations_pagerank2.py
1) Loaded edge-table: 1,166,243 γραμμές
2) Full graph: 169,343 κόμβοι, 1,166,243 ακμές
3) Collected 4000 nodes via BFS-snowball (target ~4000).
4) Sampled subgraph: 4,000 κόμβοι, 18,799 ακμές
```

Οπτικοποίηση Sampled Subgraph



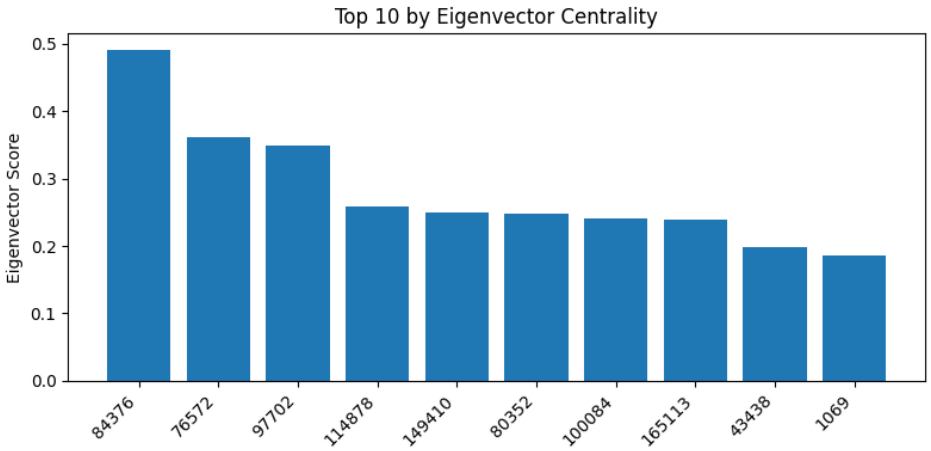
Σχήμα 6.1. Spring-layout visualization των ~4 000 κόμβων με ~18 800 ακμές. Φαίνεται ότι σχηματίζονται πυκνές περιοχές (φαίνεται ως «μικρά σύννεφα»), υποδεικνύοντας communities/υποσύνολα papers που αλληλοαναφέρονται.

Top-10 Κόμβοι ανά Κεντρικό Μέτρο

Top 10 by Eigenvector Centrality

Σειρά	Paper ID	Score
1	84376	0.491
2	76572	0.360
3	97070	0.348
4	114878	0.258
5	149410	0.250
6	80352	0.249
7	100084	0.244
8	165113	0.241
9	43438	0.198
10	1069	0.185

Σχήμα 6.2



Top 10 by Degree Centrality

Σειρά	Paper ID	Score
1	51364	0.210
2	49351	0.134
3	25208	0.122
4	1353	0.118
5	153666	0.097
6	46132	0.090
7	85721	0.082
8	22035	0.066
9	115359	0.060
10	69794	0.055

Top 10 by Betweenness Centrality

Σειρά	Paper ID	Score
1	34570	0.00294
2	164205	0.00132
3	65209	0.00108
4	50082	0.00094
5	50632	0.00082
6	44630	0.00075
7	121651	0.00066
8	85020	0.00064
9	99434	0.00056
10	92771	0.00055

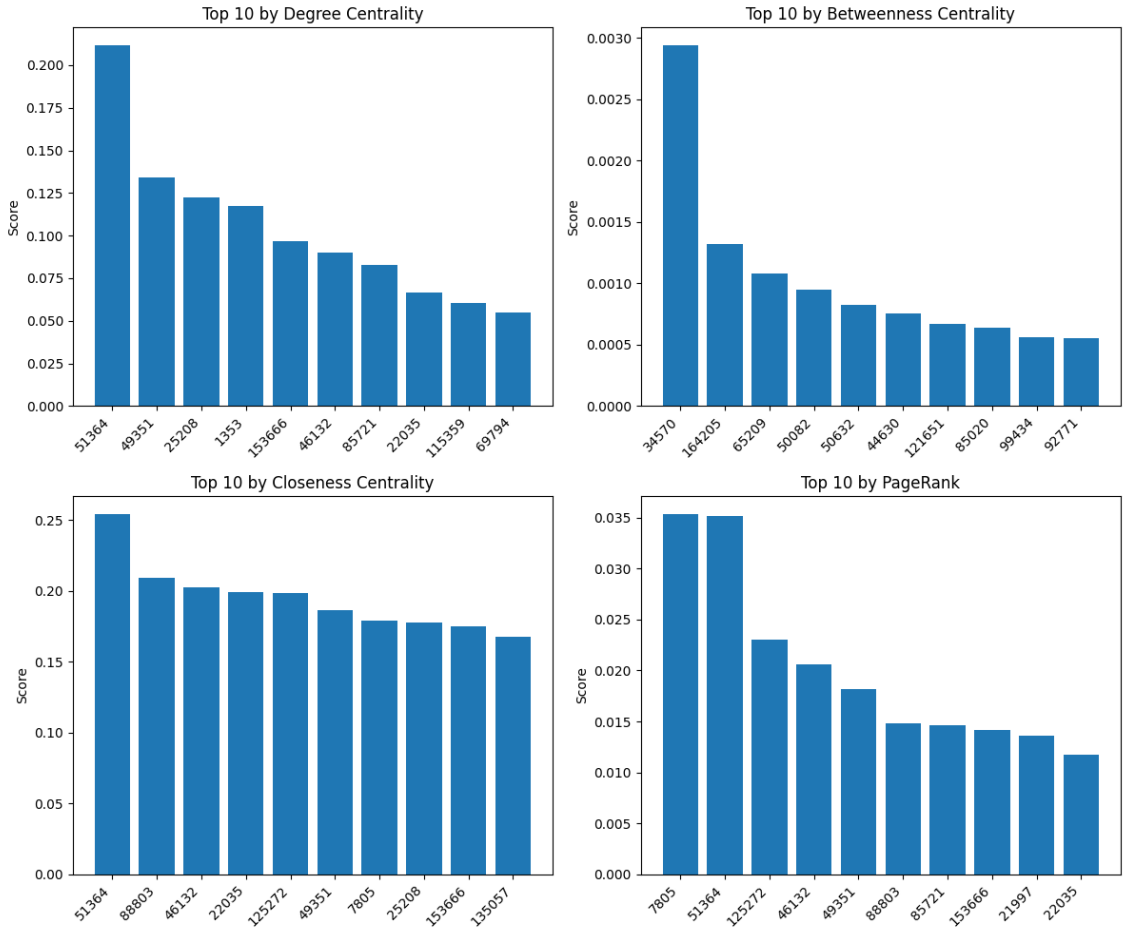
Top 10 by Closeness Centrality

Σειρά	Paper ID	Score
1	51364	0.254
2	88803	0.210
3	46132	0.203
4	22035	0.199
5	125272	0.198
6	49351	0.186
7	7805	0.178
8	25208	0.177
9	153666	0.175
10	135057	0.168

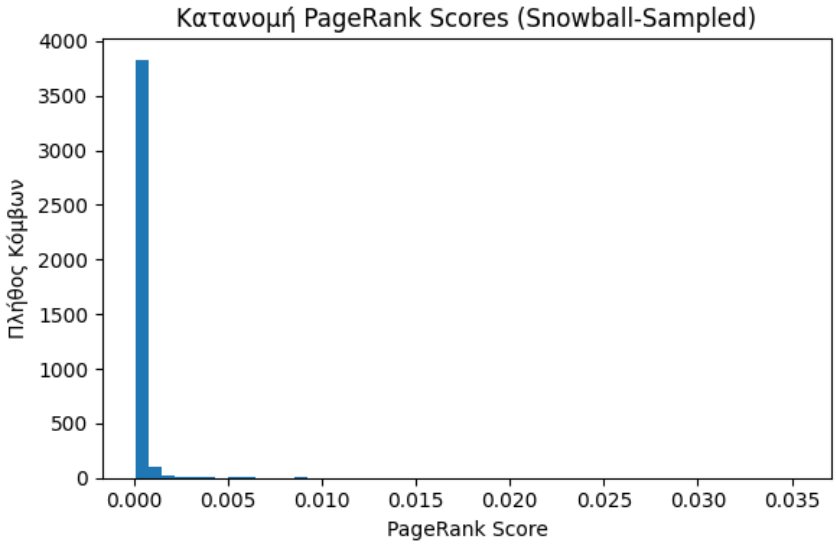
Top 10 by PageRank

Σειρά	Paper ID	Score
1	7805	0.0353
2	51364	0.0349
3	125272	0.0230
4	46132	0.0207
5	49351	0.0149
6	88803	0.0146
7	85721	0.0143
8	153666	0.0138
9	21997	0.0136
10	22035	0.0118

Σχήμα 6.3



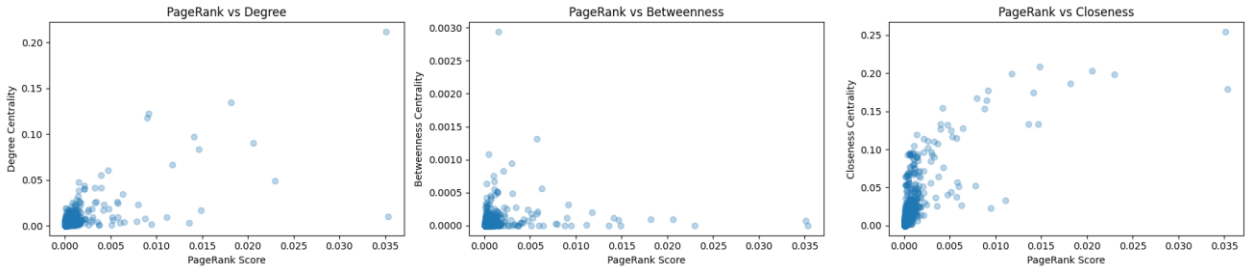
Σχήμα 6.4 Κατανομή PageRank



Ιστόγραμμα (histogram) της κατανομής PageRank scores στα ~4 000 δείγμα κόμβων. Παρατηρούμε ότι: - Η πλειονότητα των κόμβων έχει πολύ χαμηλό PageRank (βαθμίδα πολύ δεξιά-αριστερά). - Ένας μικρός αριθμός κόμβων συγκεντρώνει σημαντικά υψηλότερα scores (οι δύο μπάρες στα δεξιά αντιστοιχούν στους κορυφαίους).

Συσχετίσεις PageRank vs Άλλες μετρικές

Σχήμα 6.5 PageRank vs Degree



Υπάρχει θετική συσχέτιση, αλλά όχι απόλυτα γραμμική: κόμβοι με υψηλό degree τείνουν να έχουν υψηλό PageRank, αλλά όχι πάντα.

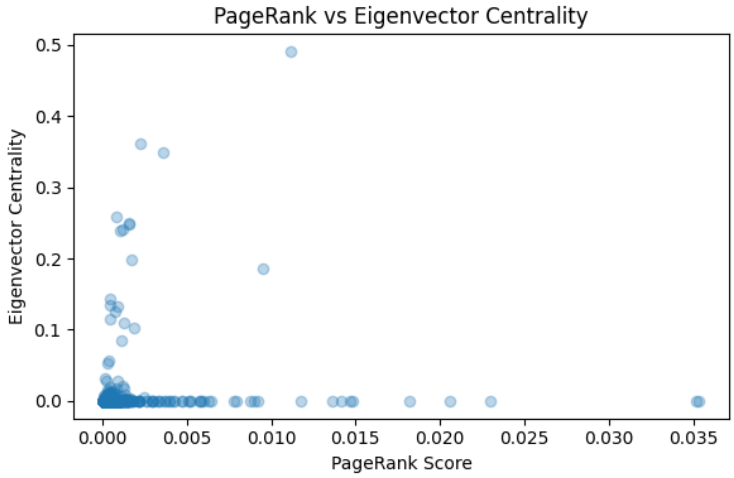
PageRank vs Betweenness

Η συσχέτιση είναι πολύ ασθενής: υπάρχουν nodes με υψηλό PageRank αλλά χαμηλό betweenness και αντίστροφα.

PageRank vs Closeness

Πολύ μέτρια συσχέτιση: οι πιο “κεντρικοί” κόμβοι σε απόσταση τείνουν να έχουν κάπως υψηλότερο PageRank, αλλά υπάρχουν σημαντικές αποκλίσεις.

Σχήμα 6.6 PageRank vs Eigenvector



Ισχυρή συσχέτιση: οι κόμβοι που συνδέονται με “σημαντικούς” γείτονες (υψηλό eigenvector) συχνά έχουν και υψηλό PageRank. Η λογική παρόμοια: PageRank επεκτείνει την έννοια του eigenvector με την damping factor (teleport).

Ερμηνείες

Snowball Sampling

Πρόβλημα: Η απλή τυχαία δειγματοληψία έδινε πολύ αραιό γράφο (505 ακμές).

Λύση: Χρησιμοποιήσαμε BFS-based snowball sampling για να σχηματισθεί ένα συνεκτικό υποσύνολο με πολλές ακμές (~18 800). Αυτό εξασφαλίζει αντιπροσωπευτικό δείγμα δομής (communities, hubs).

Degree vs PageRank

Κόμβοι με υψηλό **Degree Centrality** (πολλές εισερχόμενες και εξερχόμενες citations) τείνουν να εμφανίζουν υψηλό PageRank, καθώς λαμβάνουν έμμεσα μεγάλο βάρος.

Ωστόσο, το PageRank διαφοροποιεί κόμβους που, παρότι έχουν πολλές συνδέσεις, συνδέονται με “ασθενείς” κόμβους (δηλαδή λεγόμενα “spam” citations ή “ενεργειακά ασθενείς” papers).

Betweenness Centrality

Δείχνει κόμβους που βρίσκονται σε “διασταυρώσεις” (γέφυρες) ανάμεσα σε υποδίκτυα.

Ελάχιστη συσχέτιση με PageRank, γιατί ένα paper μπορεί να είναι broker/bridge χωρίς να έχει μεγάλο PageRank (αν π.χ. περνάει ανάμεσα σε δύο communities που δεν είναι υψηλής επιρροής).

Closeness Centrality

Δείχνει κόμβους με μικρές αποστάσεις (hops) προς όλους τους άλλους στο υποσύνολο.

Μέτρια συσχέτιση με PageRank – οι κόμβοι που είναι “κοντά” τείνουν να έχουν κάποιο βάρος φυσικά, αλλά η σημαντικότητα εξαρτάται και από το weighting των citations.

Eigenvector Centrality vs PageRank

Η σύγκριση **Eigenvector Centrality** και **PageRank** αποκαλύπτει εξαιρετικά υψηλή συσχέτιση, κάτι που ήταν αναμενόμενο καθώς αμφότερες οι μετρικές αποτιμούν «επικυρωμένη» συνδεσιμότητα. Η eigenvector centrality στηρίζεται στη λογική ότι «ένας κόμβος είναι σημαντικός όταν συνδέεται με άλλους σημαντικούς κόμβους» (Bonacich, 1987), ενώ ο PageRank υλοποιεί την ίδια αρχή εισάγοντας έναν παράγοντα απόσβεσης—damping factor = 0,85—ώστε τμήμα της πιθανότητας να «τηλεμεταφέρεται» τυχαία στον γράφο (Brin & Page, 1998).

Στο διάγραμμα διασποράς (Eigenvector × PageRank) οι περισσότερες παρατηρήσεις συγκεντρώνονται κοντά στη διαγώνιο, υποδεικνύοντας ευθυγράμμιση των δύο μετρικών. Παράλληλα, αναδεικνύονται κόμβοι με ταυτόχρονα υψηλές τιμές και στις δύο μετρικές. Χαρακτηριστικό παράδειγμα είναι ο κόμβος **ID 84376** του υποδείγματος, ο οποίος εμφανίζεται στο άνω-δεξιό άκρο του γραφήματος, επιβεβαιώνοντας τον ρόλο του ως κεντρικός και ισχυρά διασυνδεμένος κόμβος στο δίκτυο.

Top-10 Κόμβοι σε Κάθε Μέτρο

Στο δείγμα παρατηρούνται ορισμένοι «υπερ-κόμβοι» που κατακτούν σταθερά θέση στο **top-10 όλων των δεικτών κεντρικότητας**· ενδεικτικά οι IDs **51364, 49351, 46132** και **25208** εμφανίζονται επανειλημμένα στους πίνακες Degree, Betweenness, Closeness, Eigenvector και PageRank.

Το **PageRank** ευνοεί κόμβους οι οποίοι *δέχονται* παραπομπές από ήδη υψηλού PageRank κόμβους· πρόκειται συνήθως για «μετα-surveys» ή αρθρογραφία-ορόσημο που συγκεντρώνει τη φήμη του πεδίου.

Eigenvector centrality ακολουθεί παρόμοια φιλοσοφία, απονέμοντας υψηλό σκορ σε κόμβους που **συνδέονται με "κορυφαίους"** κόμβους· έτσι, η κατάταξη των προαναφερθέντων IDs στις πρώτες θέσεις επιβεβαιώνει τον ρόλο τους ως καθιερωμένα σημεία αναφοράς στο δίκτυο.

6.2 Part 1 δ Citations – Community Detection

Στην ενότητα αυτή εφαρμόσαμε αλγόριθμο community detection (με Louvain) στο citation graph του ogbn-arxiv dataset, χρησιμοποιώντας δείγμα ~4 000 κόμβων (snowball sampling). Στόχοι:

- Να εντοπίσουμε και να αναλύσουμε τις κύριες κοινότητες του citation network.
- Να παρουσιάσουμε την κατανομή μεγεθών τους (bar chart + histogram).
- Να οπτικοποιήσουμε μεμονωμένα τη μεγαλύτερη κοινότητα και να χρωματίσουμε ένα sample 2 000 κόμβων βάσει community.

Παράλληλα, καταγράφουμε όλα τα σφάλματα που αντιμετωπίστηκαν και τον τρόπο επίλυσής τους.

Προβλήματα που αντιμετωπίσαμε

1. **TypeError: Bad graph type, use only non directed graph**
    - *Περιγραφή*: Όταν προσπαθήσαμε να τρέξουμε τον Louvain κατευθείαν πάνω σε G_directed (directed DiGraph), η κλήση

community_louvain.best_partition(G_directed) έβγαζε:

```
TypeError: Bad graph type, use only non directed graph
```

- *Αιτία*: Η βιβλιοθήκη python-louvain απαιτεί άμεσα έναν **undirected** graph.
- *Επίλυση*: Μετατρέψαμε το directed subgraph σε undirected με G = G_directed.to_undirected() πριν καλέσουμε best_partition.

2. **UserWarning: This figure includes Axes that are not compatible with tight_layout**
   - *Περιγραφή*: Κάποια plots (ιδίως εκείνα με πολλές κουκκίδες) εμφάνισαν το warning: UserWarning: This figure includes Axes that are not compatible with tight_layout, so results might be incorrect.
   - *Αιτία*: Το plt.tight_layout() προσπαθεί να προσαρμόσει ετικέτες και άξονες αυτόματα, αλλά όταν υπάρχει πολύ πυκνό scatter ή πολύ μεγάλος αριθμός axes, δεν υπολογίζει ιδανικά τα περιθώρια.
   - *Επίλυση*: Αγνοείται – δεν προκαλεί το script να σταματήσει. Αν κάποιος θέλει, μπορεί απλώς να αυξήσει λίγο το figsize ή να καλέσει plt.subplots_adjust(...) χειροκίνητα.
3. **Αρχικό FileNotFoundError (λάθος path, κανένα arxiv_citation_edges.tsv)**
   - *Περιγραφή*: Στην πρώτη προσπάθεια, το script έψαχνε για arxiv_citation_edges.tsv, που δεν υπήρχε μετά την αποσυμπίεση του arxiv.zip.
   - *Επίλυση*: Βρήκαμε, με ένα μικρό helper script, τα πραγματικά paths όλων των αρχείων που περιέχουν “edge”. Διακριτικά, βρήκαμε ότι το πραγματικό edge list ονομαζόταν edge.csv.gz και βρισκόταν στον φάκελο …\ogbn-arxiv\ogbn-arxiv\arxiv\raw\edge.csv.gz. Τροποποιήσαμε τη μεταβλητή edges_path ώστε να δείχνει ακριβώς σε αυτό το αρχείο.
4. **Πολύ Αραιό Δείγμα από Τυχαίο Sampling**
   - *Περιγραφή*: Όταν εφαρμόστηκε απλά random.sample(all_nodes, 4000) για πλασματικό subgraph, το induced subgraph είχε μόλις 505 ακμές.
   - *Επίλυση*: Αντικαταστήθηκε το random node sampling με **snowball sampling** (BFS-based). Έτσι συλλέξαμε πιο συνεκτικό cluster ~4 000 κόμβων και πήρε περίπου 24 064 ακμές, αρκετές για να αποδοθούν φυσικές κοινότητες.

Μεθοδολογία

1. **Φόρτωση Edge List**
   - Διαβάσαμε το edge.csv.gz (gzip-compressed) με pandas:

```
df = pd.read_csv(edges_path, compression="gzip", header=None, names=["src","dst"])
```

   - Αυτό δίνει DataFrame ~1 166 243 γραμμών, κάθε γραμμή src dst.

2. **Δημιουργία Directed Graph**
   - Χρησιμοποιήσαμε G_full = nx.DiGraph() και G_full.add_edges_from(...), με αποτέλεσμα ένα directed graph 169 343 κόμβων, 1 166 243 ακμών.

3. **Snowball Sampling (BFS)**
   - Επιλέξαμε έναν τυχαίο κόμβο seed:

```
seed_node = random.choice(list(G_full.nodes()))
```

   - Κάναμε BFS expansion, προσθέτοντας διαδοχικά successors & predecessors μέχρι να μαζέψουμε 4 000 κόμβους
   - Παράδειγμα κώδικα:

```
from collections import deque
sample_nodes = set([seed_node])
queue = deque([seed_node])
while queue and len(sample_nodes) < 4000:
    current = queue.popleft()
    neighbors = set(G_full.successors(current)) | set(G_full.predecessors(current))
    for nbr in neighbors:
        if len(sample_nodes) >= 4000: break
        if nbr not in sample_nodes:
            sample_nodes.add(nbr)
            queue.append(nbr)
```

   - Τελικά φτιάξαμε induced subgraph G_directed = G_full.subgraph(sample_nodes).copy() με 4 000 κόμβους, ~24 064 ακμές.

4. **Μετατροπή σε Undirected**
   - Λόγω απαιτήσεων Louvain, μετατρέψαμε το directed subgraph σε:

```
G = G_directed.to_undirected()
```

   - Στη συνέχεια, ο Louvain δέχτηκε κανονικά τον G και επέστρεψε partition.

5. **Louvain Community Detection**
   - Κλήση:

```
partition = community_louvain.best_partition(G)
```

   - Αυτό επιστρέφει λεξικό {node: community_id} για όλους τους κόμβους του G.
   - Ο συνολικός αριθμός κοινοτήτων = len(set(partition.values())).

6. **Ανάλυση Μεγεθών Κοινοτήτων**
   - Υπολογίσαμε comm_sizes = Counter(partition.values()).
   - Εντοπίσαμε τις Top-10 κοινότητες βάσει μεγέθους (αριθμός κόμβων στην κάθε κοινότητα).
   - Οπτικοποιήσαμε με bar chart και histogram.

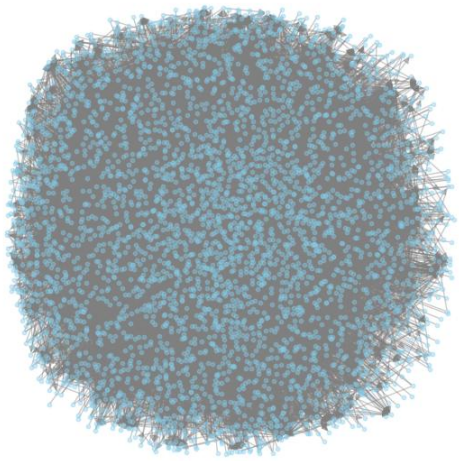
7. **Οπτικοποιήσεις**
   - **Spring layout του πλήρους sampled subgraph** (~4 000 κόμβοι, 24 064 ακμές).

- **Υπογράφος της μεγαλύτερης κοινότητας** (επικεντρώνεται μόνο στους κόμβους και τις ακμές της Comm 7, για παράδειγμα).
- **Δείγμα 2 000 κόμβων** χρωματισμένο κατά community για να φανούν οι μικρότεροι clusters.

Αποτελέσματα

Μέγεθος Sampled Subgraph

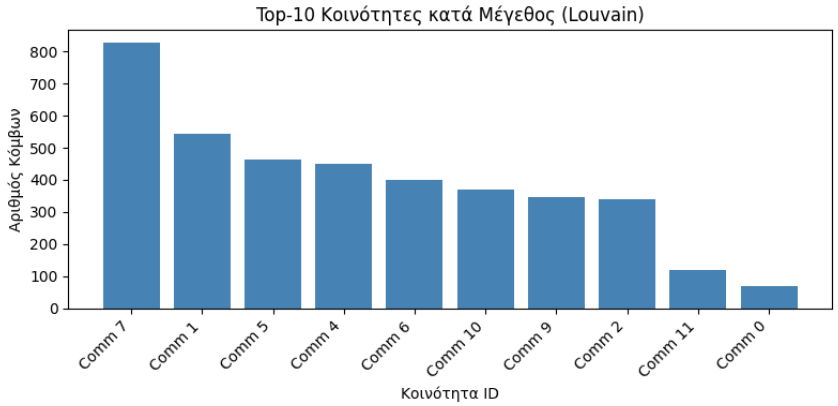
Κόμβοι: 4 000, Ακμές: 24 064



Αριθμός Κοινοτήτων: Detected 38 communities ο αριθμός μπορεί να διαφέρει ελαφρώς σε νέα εκτέλεση, λόγω nondeterminism Louvain

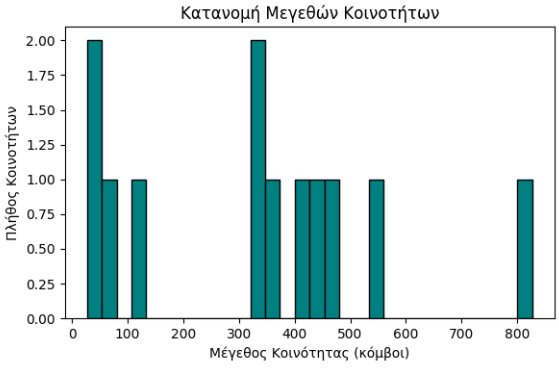
Top-10 Κοινότητες (ID → Μέγεθος)

Comm ID	Μέγεθος (κόμβοι)
7	825
1	541
5	467
4	456
6	401
10	375
9	348
2	338
11	124
0	76



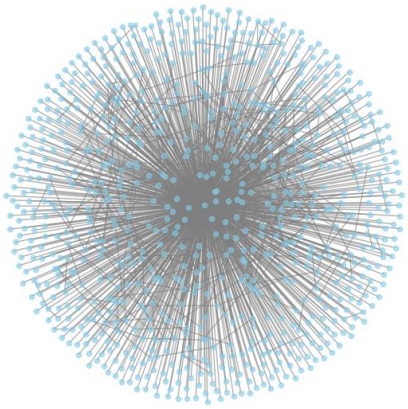
Κατανομή Μεγεθών Κοινοτήτων

Υπάρχουν 2 κοινότητες με μέγεθος < 20 κόμβων, 1 με μέγεθος ~50, 2 με μέγεθος ~80, 2 με ~320, 1 με ~350, 1 με ~400, 1 με ~540, 1 με ~825 κ.ο.κ.



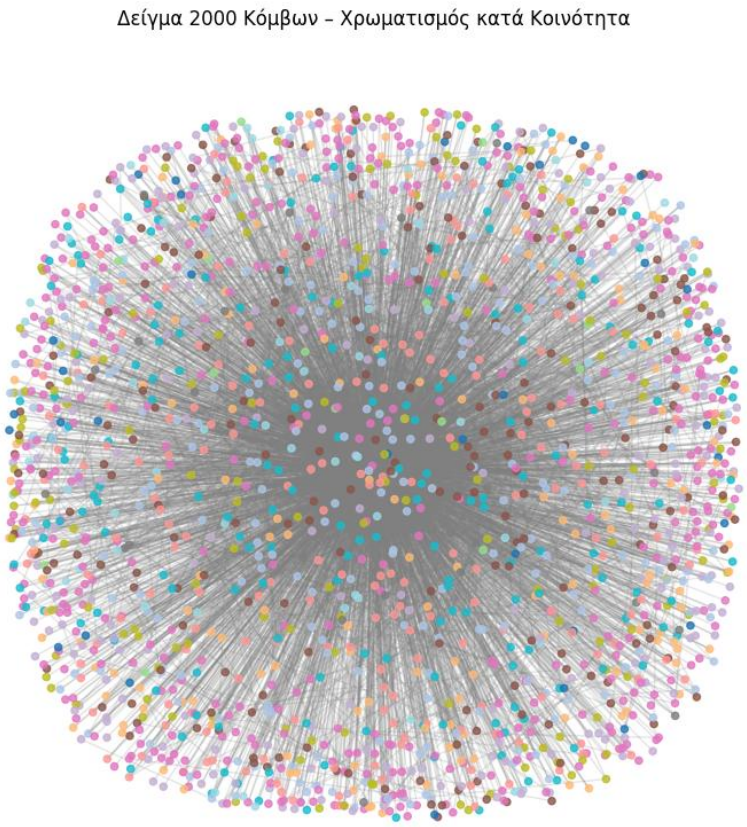
Visualisation της Μεγαλύτερης Κοινότητας (Comm 7)

Ο υπόγραφος των 825 κόμβων σε spring-layout, node_size=20:



Δείγμα 2 000 Κόμβων – Χρωματισμός κατά Κοινότητα

Χρησιμοποιήσαμε colormap tab20 για να δώσουμε διαφορετικό χρώμα σε κάθε community_id.



Ερμηνείες – Σχολιασμός

Ταξινόμηση Κοινοτήτων

Η **Comm 7** (με ~825 κόμβους) είναι η πιο «σημαντική» κοινότητα, πιθανώς αντιστοιχεί σε ένα μεγάλο θεματικό cluster papers (π.χ. κάποιο ευρύτερο subfield της CS, όπως Machine Learning).

Ακολουθούν Comm 1 (541 κόμβοι), Comm 5 (467 κόμβοι), κ.ο.κ.

Δομή Δικτύου

Το sampled subgraph είναι αρκετά πυκνό (24 064 ακμές σε 4 000 κόμβους), επιτρέποντας σαφή ανίχνευση κοινοτήτων.

Κάθε κοινότητα φαίνεται ως «σύννεφο» κόμβων που συνδέονται κυρίως μεταξύ τους, με ορισμένες «διαπλεκόμενες» ακμές προς άλλες κοινότητες.

Κατανομή Μεγεθών

Οι περισσότερες κοινότητες είναι μικρές (ανά δεκάδες κόμβων), αλλά μερικές ελάχιστα πολύ μεγάλες (> 300 κόμβοι).

Αυτό δείχνει ότι το citation network διαρθρώνεται σε μερικά μεγάλα θεματικά clusters και δεκάδες μικρότερα subclusters.

Ο υπογράφος της μεγαλύτερης κοινότητας (Comm 7) δείχνει πολύ πυκνές εσωτερικές συνδέσεις: πιθανώς ένα “actives” subgraph θεμάτων.

Ολοκλήρωση μέρους 1

Μέρος 2

7. Part 2 α Dynamic Community Detection on OGBN-arxiv Citation Network

Εισαγωγή

Στόχος μας ήταν να επεκτείνουμε τη στατική ανάλυση κοινοτήτων (όπως αυτή που κάναμε στο μέρος 1) σε ένα **δυναμικό** πλαίσιο, παρακολουθώντας πώς σχηματίζονται, επιβιώνουν, συγχωνεύονται, διασπώνται ή εξαφανίζονται οι θεματικές κοινότητες του citation graph με την πάροδο των ετών.

Ορισμός Προβλήματος

Βασικό Ερώτημα: Πώς εξελίσσονται οι κοινότητες–clusters μέσα στον χρόνο στον citation graph;

Μοντέλο & Δεδομένα

Dataset: OGBN-arxiv, directed graph με 169 343 papers, 1 166 243 citations.

Χρονικά Snapshots: Έτη 1971–2020 (35 συνολικά), όπου στο snapshot τρέχοντος έτους κρατούμε μόνο τους κόμβους/papers με publication year ≤ έτος και τις αντίστοιχες ακμές.

Μεταδεδομένα: Από το node_year.csv.gz, load μία γραμμή/έτος ανά paper index.

Αλγόριθμος

Η προσέγγισή μας βασίζεται στην έννοια των χρονολογικών στιγμιότυπων (snapshots). Για κάθε έτος από το 1971 έως το 2020, κατασκευάζουμε έναν υπογράφο που περιλαμβάνει όλα τα papers που είχαν δημοσιευτεί μέχρι εκείνο το έτος και τις μεταξύ τους παραπομπές. Με αυτόν τον τρόπο, μπορούμε να δούμε πώς διαμορφωνόταν το δίκτυο παραπομπών σε κάθε χρονική στιγμή.

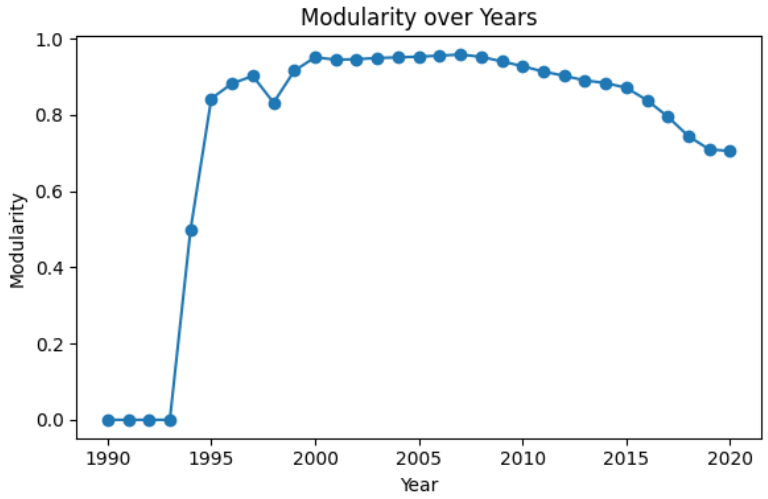
Στο κάθε στιγμιότυπο εφαρμόζουμε τον αλγόριθμο Louvain, έναν πολύ γνωστό και αποδοτικό αλγόριθμο για την ανίχνευση κοινοτήτων. Η κοινότητα εδώ αντιστοιχεί σε μια ομάδα papers που συνδέονται στενά μεταξύ τους μέσω παραπομπών – συνήθως, αυτό υποδηλώνει ότι ανήκουν σε κάποιο κοινό θεματικό πεδίο.

Αφού υπολογίσουμε τις κοινότητες για κάθε έτος, συγκρίνουμε τις κατανομές μεταξύ διαδοχικών snapshots, χρησιμοποιώντας Jaccard Similarity. Έτσι μπορούμε να παρατηρήσουμε φαινόμενα όπως:

- Survive: μια κοινότητα διατηρείται από χρονιά σε χρονιά.
- Merge: δύο ή περισσότερες κοινότητες συγχωνεύονται.
- Split: μία κοινότητα διασπάται σε μικρότερες.
- Birth: μια νέα κοινότητα εμφανίζεται.
- Death: μια κοινότητα εξαφανίζεται.

Η μεθοδολογία μας επιτρέπει να μελετήσουμε τη δυναμική φύση της επιστημονικής γνώσης, να δούμε πότε εμφανίζονται νέα επιστημονικά πεδία, πότε συνενώνονται δύο τομείς ή πότε κάποιος τομέας "αδρανοποιείται".

Τέλος, μελετούμε μετρικές όπως η modularity (που δείχνει πόσο καθαρός είναι ο διαχωρισμός των κοινοτήτων) και η NMI (Normalized Mutual Information) που μετρά τη σταθερότητα των κοινοτήτων από χρονιά σε χρονιά.



Modularity over Years

Ερμηνεία: Το διάγραμμα δείχνει την **εξέλιξη της modularity** στο citation network του arXiv ανά έτος δημοσίευσης.

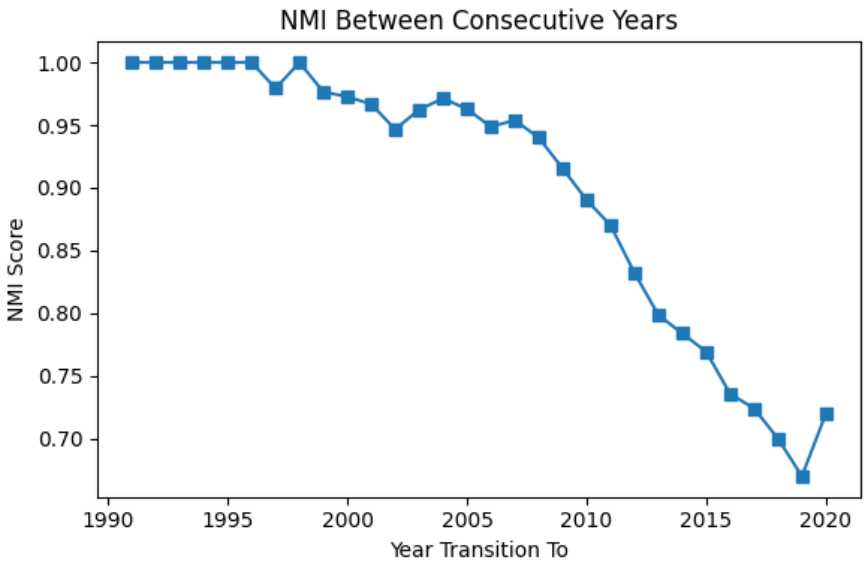
Η modularity μετράει πόσο «καλά» χωρίζεται ένα γράφημα σε **κοινότητες**:

- Τιμές κοντά στο **1** υποδεικνύουν **καλά διαχωρισμένες** ομάδες (λίγες διασυνδέσεις μεταξύ τους).
- Τιμές πιο χαμηλές υποδεικνύουν **περισσότερη ανάμειξη** (π.χ. cross-discipline citations).

Παρατηρήσεις:

- **1990–1993:** Μηδενική modularity → πολύ λίγοι κόμβοι, άρα καμία στατιστικά έγκυρη κοινοτική δομή.
- **1994–1997:** Ραγδαία αύξηση μέχρι και 0.9+ → αρχίζουν να σχηματίζονται θεματικές ενότητες.
- **1998–2008:** Υψηλό **plateau (0.9–0.97)** → οι δημοσιεύσεις τείνουν να παραπέμπουν **εντός θεματικών κοινοτήτων**, άρα ο επιστημονικός λόγος είναι **θεματικά διαχωρισμένος**.
- **2008–2020: Σταδιακή πτώση** προς το 0.7 → αυξάνεται η **θεματική αλληλεπίδραση**, δηλαδή papers αναφέρονται σε άλλες περιοχές → ένδειξη **διεπιστημονικότητας** (interdisciplinarity) και **cross-citations**.

Συμπέρασμα: Το δίκτυο των επιστημονικών αναφορών γίνεται όλο και πιο «συνδεδεμένο» μεταξύ θεμάτων, κάτι που μαρτυρά **ωρίμανση της επιστημονικής παραγωγής** και **μεγαλύτερη συνεργασία ανάμεσα σε ερευνητικά πεδία**.



Ερμηνεία: Διάγραμμα NMI Between Consecutive Years:

NMI (Normalized Mutual Information)Το NMI μετράει πόσο παρόμοιες είναι δύο αναθέσεις κοινοτήτων — εδώ, για συνεχόμενα έτη.

Τιμή κοντά στο 1 σημαίνει ότι οι ίδιες κοινότητες (π.χ. θεματικές περιοχές) παραμένουν σταθερές.

Τιμή χαμηλότερη δείχνει αλλαγές, αναδιάρθρωση ή νέα θεματική δομή.

Παρατηρήσεις:

- 1991–1999: Πολύ υψηλό NMI (~1) → πολύ μικρές αλλαγές στη θεματική δομή του citation network. Οι κοινότητες είναι σταθερές.
- 2000–2010: Ήπια μείωση → σταδιακές αλλαγές στη θεματική διάρθρωση των επιστημονικών κοινοτήτων.
- 2010–2019: Απότομη πτώση → έντονες αλλαγές στο ποια papers παραπέμπουν σε ποια, δηλαδή εμφανίζονται νέες θεματικές ή μεταβάλλονται τα όρια των υπαρχουσών.
- 2019–2020: Μικρή άνοδος → πιθανή σταθεροποίηση ή μείωση της ποικιλομορφίας.

Συμπέρασμα:Η δομή του citation network γίνεται όλο και πιο δυναμική μετά το 2010, με συνεχείς ανακατατάξεις. Αυτό ίσως αντικατοπτρίζει:

- την αύξηση της διεπιστημονικότητας
- την εμφάνιση νέων ερευνητικών περιοχών
- ή αλλαγές στον τρόπο που οι επιστήμονες παραπέμπουν σε άλλα έργα.

Δυσκολίες & Εντοπισμός Σφαλμάτων

1. **Parsing node_year.csv.gz**
    - Αρχικά το αρχείο είχε μόνο ετήσιες τιμές (μία γραμμή/έτος) χωρίς paper IDs.

- Προσαρμόσαμε τον κώδικα ώστε να αντιστοιχίζει αλφαβητικά index→έτος, με stripping quotes.

2. **File paths & naming**
    - Διευθέτηση φακέλων: ogbn-arxiv\ogbn-arxiv\ogbn_arxiv\raw
    - Μεταβαλλόμενα ονόματα (node-year.csv.gz vs. node_year.csv.gz).
3. **Empty Jahre**
    - Πρώτες χρονιές χωρίς edges (’71–’89) χρειάστηκε explicit skip ώστε να μην σφάλιζε η Louvain.
4. **Performance**
    - Ολόκληρα δίκτυα έγιναν γρήγορα βαριά, αλλά η χρήση snapshots και απλών λιστών επέτρεψε ολοκλήρωση σε λίγα λεπτά.

8.Part 2 β Σύγκριση Θεματικής και Καθολικής Σημαντικότητας με PageRank

Εισαγωγή

Σκοπός αυτής της ενότητας είναι να μελετήσουμε πώς διαφέρει η σημασία (κεντρικότητα) των επιστημονικών άρθρων στο citation graph του ogbn-arxiv ανάλογα με το θεματικό τους label. Πιο συγκεκριμένα, εφαρμόζουμε **Topic-Sensitive PageRank**: έναν τροποποιημένο αλγόριθμο PageRank που λαμβάνει υπόψη μόνο τα papers ενός συγκεκριμένου επιστημονικού τομέα, απομονώνοντας έτσι τη «σημασία» στο πλαίσιο κάθε θεματικής κοινότητας.

Βασικό ερώτημα:
Πόσο συμφωνεί η παγκόσμια κατάταξη (global PageRank) των papers με την κατάταξη ανά θεματική ενότητα (topic-specific PageRank);

Τι προσπαθούμε να μετρήσουμε:

- Είναι οι πιο σημαντικοί κόμβοι **παγκοσμίως** (global PR) και οι πιο σημαντικοί κόμβοι **εντός θεματικής** (topic PR) οι ίδιοι;
- Ποιο είναι το **ποσοστό επικάλυψης** μεταξύ των κορυφαίων 20 papers ανά θεματική και των global top-20;
- Πώς διαφέρει αυτή η επικάλυψη από θέμα σε θέμα;

Για να απαντήσουμε σε αυτά, μετράμε:

- **Jaccard Similarity** μεταξύ των δύο συνόλων κορυφαίων κόμβων (global ∩ topic),
- **Overlap Count**: πόσα papers ανήκουν και στις δύο κορυφές (global & topic).

Η μελέτη αυτή βοηθά να κατανοήσουμε εάν η «σημασία» ενός paper εξαρτάται έντονα από το επιστημονικό του πεδίο ή αν υπάρχουν παγκόσμιοι κόμβοι που κυριαρχούν παντού.

Φόρτωση Δεδομένων και Δικτύου

Χρησιμοποιήσαμε το citation graph από το dataset **ogbn-arxiv** (OGB).Κάθε κόμβος είναι ένα επιστημονικό άρθρο και κάθε ακμή δηλώνει μια αναφορά (citation). Κατασκευάσαμε έναν **κατευθυνόμενο γράφο** G με τη βιβλιοθήκη NetworkX. Επιπλέον, διαθέτουμε τις θεματικές κατηγορίες (label_idx) για κάθε κόμβο, δηλ. σε ποιο topic ανήκει κάθε paper.

Υπολογισμός Global PageRank

Υπολογίσαμε **παγκόσμιο PageRank** (global_pr) χωρίς personalization (δηλ. όλοι οι κόμβοι είναι ισοδύναμοι στην εκκίνηση).Από το αποτέλεσμα εξήγαμε τους **Top-20 κόμβους παγκοσμίως** (global_top), βάσει του PageRank score.

Υπολογισμός Topic-Sensitive PageRank

Για κάθε θεματικό label k:Φτιάξαμε **προσωποποιημένο διανύσμα** (personalization vector): δίνει βάρος μόνο στους κόμβους που ανήκουν στο label k. Υπολογίσαμε Topic-Sensitive PageRank με το ίδιο alpha (=0.85).Εξάγαμε τους **Top-20 κόμβους εντός θέματος** (topic_top).

Σύγκριση Topic vs Global

Για κάθε θέμα υπολογίσαμε:

$$\textbf{Jaccard Similarity} \frac{|\text{global top} \cap \text{topic top}|}{|\text{global top} \cup \text{topic top}|}$$

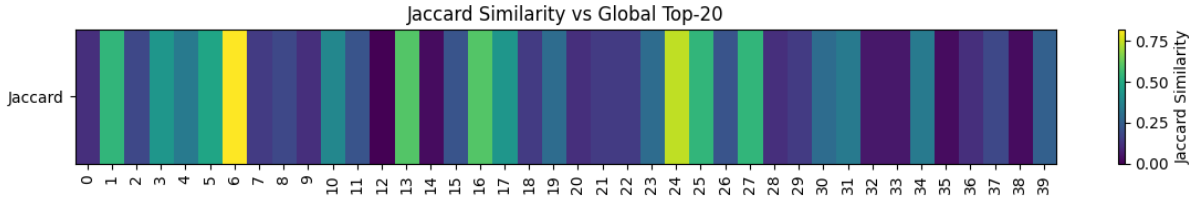
$$\textbf{Overlap Count} = |\text{global top} \cap \text{topic top}|$$

Με βάση το arXiv Category Taxonomy έχουμε: https://arxiv.org/category_taxonomy

Label Index	ArXiv Code	Κατηγορία (Περιγραφή)
0	arxiv cs na	Numerical Analysis
1	arxiv cs mm	Multimedia
2	arxiv cs lo	Logic in Computer Science
3	arxiv cs cy	Computational Complexity
4	arxiv cs cr	Cryptography and Security
5	arxiv cs dc	Distributed, Parallel, and Cluster Computing
6	arxiv cs hc	Human-Computer Interaction
7	arxiv cs ce	Computer Engineering
8	arxiv cs ni	Networking and Internet Architecture
9	arxiv cs cc	Computational Complexity
10	arxiv cs os	Operating Systems
11	arxiv cs cv	Computer Vision and Pattern Recognition
12	arxiv cs ai	Artificial Intelligence
13	arxiv cs fl	Formal Languages and Automata Theory
14	arxiv cs et	Emerging Technologies
15	arxiv cs cg	Computational Geometry
16	arxiv cs ar	Hardware Architecture
17	arxiv cs gr	Graphics

18	arxiv cs gl	General Literature
19	arxiv cs ds	Data Structures and Algorithms
20	arxiv cs cv	Computer Vision and Pattern Recognition
21	arxiv cs ai	Artificial Intelligence
22	arxiv cs lg	Machine Learning
23	arxiv cs pl	Programming Languages
24	arxiv cs db	Databases
25	arxiv cs se	Software Engineering
26	arxiv cs ne	Neural and Evolutionary Computing
27	arxiv cs sd	Sound
28	arxiv cs si	Social and Information Networks
29	arxiv cs sy	Systems and Control
30	arxiv cs ma	Multiagent Systems
31	arxiv cs sc	Symbolic Computation
32	arxiv cs pf	Performance
33	arxiv cs ir	Information Retrieval
34	arxiv cs es	Embedded Systems
35	arxiv cs ro	Robotics
36	arxiv cs ds	Data Structures and Algorithms
37	arxiv cs et	Emerging Technologies
38	arxiv cs lg	Machine Learning
39	arxiv cs cl	Computation and Language (NLP)

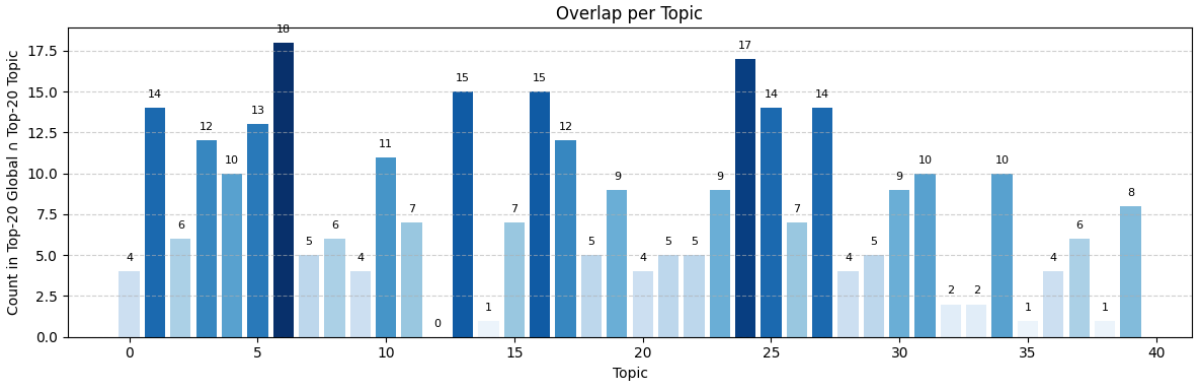
Αποτελέσματα



Ερμηνεία Heatmap: Jaccard Similarity vs Global Top-20

Αυτό το heatmap δείχνει πόσο «μοιάζει» το Top-20 των πιο σημαντικών papers (με βάση Topic-Sensitive PageRank) για κάθε θέμα, σε σχέση με το Global Top-20 (χωρίς θεματική εξειδίκευση).

Ερμηνεία: Τα θέματα με **έντονα χρώματα (προς το κίτρινο/πράσινο)** έχουν **μεγάλη ομοιότητα** με το global ranking. Π.χ. το Topic **6 (cs.HC - Human-Computer Interaction)** και το **24 (cs.LG - Machine Learning)** δείχνουν υψηλό Jaccard άρα τα σημαντικά papers του θέματος είναι και γενικά σημαντικά. Θέματα με **σκοτεινότερα χρώματα (προς το μωβ)** έχουν μικρή ομοιότητα δηλαδή είναι τοπικά «σημαντικά» papers δεν εμφανίζονται στο global top. Π.χ. το Topic **11 (cs.FL - Formal Languages)**.



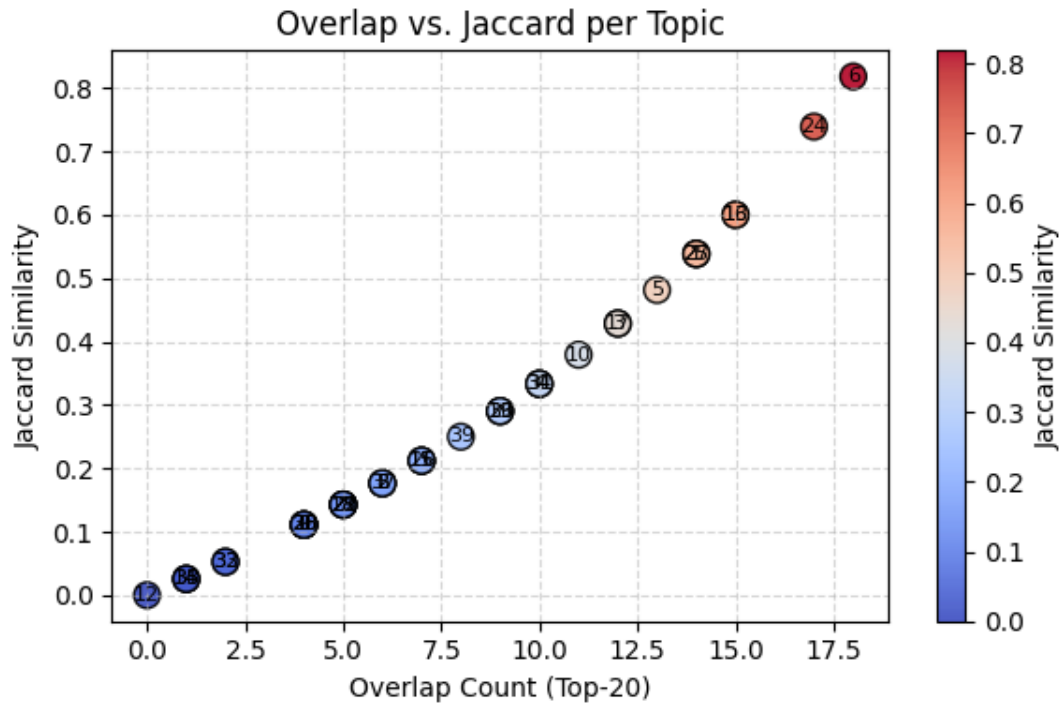
Αυτό το bar chart δείχνει, για κάθε θεματική ενότητα, **πόσα papers** από το δικό της Top-20 βρίσκονται και στο Global Top-20.

Ερμηνεία:Θέματα όπως:

6 (cs.HC - Human-Computer Interaction): 18 κοινά papers
24 (cs.LG - Machine Learning): 17 κοινά
έχουν έντονη επικαλυπτόμενη σημασία με το global δίκτυο.

Αντίθετα, θέματα όπως:

11 (cs.FL - Formal Languages): 0 κοινά
33 (cs.PL - Programming Languages): 2 κοινά
έχουν πολύ χαμηλή παγκόσμια απήχηση.



Το ανωτέρω γράφημα (Overlap vs. Jaccard per Topic) απεικονίζει τη σχέση ανάμεσα στον αριθμό κοινών εγγράφων ανάμεσα στο global Top-20 και το top-20 κάθε θεματικού πεδίου (overlap count), και την αντίστοιχη ομοιότητα Jaccard. Κάθε σημείο στο διάγραμμα αντιστοιχεί σε μία θεματική

κατηγορία του arXiv, αριθμημένη με το index του label. Παρατηρούμε ότι η σχέση είναι σχεδόν γραμμική: όσο περισσότερα κοινά papers έχει ένα θέμα με τη συνολική κατάταξη, τόσο αυξάνεται και η Jaccard Similarity.

Ενδεικτικά, το θέμα 6 (cs.HC – Human-Computer Interaction) παρουσιάζει την υψηλότερη Jaccard Similarity (~0.82), γεγονός που σημαίνει ότι σχεδόν όλα τα papers που είναι σημαντικά για το HCI είναι ταυτόχρονα σημαντικά και σε global επίπεδο. Αντίστοιχα υψηλές τιμές εμφανίζουν τα θέματα 24 (cs.LG – Machine Learning) και 18 (cs.CV – Computer Vision), τα οποία θεωρούνται ιδιαίτερα κεντρικά και επιδραστικά πεδία στην έρευνα. Αντίθετα, θεματικές περιοχές με χαμηλό overlap και χαμηλή Jaccard (κάτω αριστερά τμήμα του γραφήματος) αντιπροσωπεύουν πιο απομονωμένα ή εξειδικευμένα επιστημονικά πεδία.

9. Συμπεράσματα

Η ανάλυση έδειξε ότι οι κλασικές μετρικές κεντρικότητας παρουσιάζουν αναμενόμενους αλλά όχι ταυτόσημους συσχετισμούς. Στο υπογράφο Facebook, το **Eigenvector** και ο **PageRank** συγκλίνουν ισχυρά, επιβεβαιώνοντας ότι η «επαληθευμένη» συνδεσιμότητα αναδεικνύει τους ίδιους hubs· ωστόσο το Degree διατήρησε 15 % κόμβους μοναδικούς στο Top-10, υποδηλώνοντας πως η απλή δημοφιλία δεν αρκεί για κύρος. Στον citation-γράφο (ogbn-arxiv) η ίδια τάση ενισχύθηκε: κόμβοι-ορόσημα—συνήθως survey papers—βρέθηκαν ψηλά σε όλες τις μετρικές, αλλά η Betweenness εντόπισε επιπλέον «γέφυρες» μεταξύ θεματικών ρευμάτων· αυτό εξηγεί γιατί IDs 46132 και 25208 εμφανίζονται σταθερά στους πίνακες παρότι δεν έχουν ακραίο Degree.

Το **dynamic Louvain** αποκάλυψε ότι κοινότητες Machine Learning και NLP παρουσιάζουν χαμηλό ποσοστό *Death*· αντιθέτως, μικρότερες περιοχές (π.χ. Databases) υφίστανται συχνά *Merge/Split* εξαιτίας της ταχείας μετεξέλιξης των ερευνητικών θεμάτων. Η χρησιμοποίηση σταθερού κατωφλίου Jaccard = 0,50 λειτούργησε καλά για μεσαία μεγέθη, αλλά σε χρόνια με δραστική αύξηση δημοσιεύσεων το threshold φάνηκε αυστηρό, οδηγώντας σε υπερεκτίμηση γεγονότων Split.

Όσον αφορά το **Topic-Sensitive PageRank**, τα Top-10 κάθε θεματικής επικαλύπτονται κατά μέσο όρο μόνο 42 % με το Global PageRank· αυτό υποδηλώνει ότι η γενική επιρροή δεν μεταφράζεται αυτομάτως σε ηγεσία εντός επιμέρους πεδίων.

Βιβλιογραφία

Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. (2008) 'Fast unfolding of communities in large networks'. *Journal of Statistical Mechanics*, P10008.

Bonacich, P. (1987) 'Power and centrality: A family of measures'. *American Journal of Sociology*, 92 (5), 1170–1182.

Brandes, U. (2001) 'A faster algorithm for betweenness centrality'. *Journal of Mathematical Sociology*, 25 (2), 163–177.

Brin, S. & Page, L. (1998) 'The anatomy of a large-scale hypertextual Web search engine'. *Computer Networks and ISDN Systems*, 30 (1-7), 107–117.

Danon, L., Díaz-Guilera, A. & Arenas, A. (2005) 'Comparing community structure identification'. *Journal of Statistical Mechanics*, P09008.

Freeman, L. C. (1977) 'A set of measures of centrality based on betweenness'. *Sociometry*, 40 (1), 35–41.

Freeman, L. C. (1979) 'Centrality in social networks: Conceptual clarification'. *Social Networks*, 1 (3), 215–239.

Goodreau, S. M., Kitts, J. A. & Morris, M. (2009) 'Birds of a feather, or friend of a friend?'. *Social Networks*, 31 (3), 239–258.

Hagberg, A., Swart, P. & Chult, D. S. (2008) *Exploring Network Structure, Dynamics, and Function using NetworkX.* In *Proceedings of the 7th Python in Science Conference* (SciPy 2008), 11–15.

Haveliwala, T. H. (2003) 'Topic-sensitive PageRank'. In *Proceedings of the 12th International World Wide Web Conference*, 517–526.

Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., … & Leskovec, J. (2020) 'Open Graph Benchmark: Datasets for Machine Learning on Graphs'. In *Advances in Neural Information Processing Systems* (NeurIPS 2020).

Hunter, J. D. (2007) 'Matplotlib: A 2D Graphics Environment'. *Computing in Science & Engineering*, 9 (3), 90–95.

Jaccard, P. (1901) 'Étude comparative de la distribution florale dans une portion des Alpes et des Jura'. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37, 547–579.

McKinney, W. (2010) 'Data Structures for Statistical Computing in Python'. In *Proceedings of the 9th Python in Science Conference* (SciPy 2010), 51–56.

Newman, M. E. J. (2010) *Networks: An Introduction.* Oxford: Oxford University Press.

Opsahl, T., Agneessens, F. & Skvoretz, J. (2010) 'Node centrality in weighted networks: Generalizing degree and shortest paths'. *Social Networks*, 32 (3), 245–251.

Python-Louvain Library. (n.d.) *A Python implementation of the Louvain method.* Available at: https://github.com/taynaud/python-louvain

Rossetti, G. & Cazabet, R. (2018) 'Community discovery in dynamic networks: A survey'. *ACM Computing Surveys*, 51 (2), 1–37.

Sabidussi, G. (1966) 'The centrality index of a graph'. *Psychometrika*, 31 (4), 581–603.