УДК 004.67

І. В. Балагура, Д. В. Ланде, І. В. Горбов
Інститут проблем реєстрації інформації НАН України
вул. М. Шпака, 2, 03113 Київ, Україна

Дослідження параметрів важливості вузлів у мережах співавторів

Досліджено мережу співавторства в галузях «Інформаційна та обчислювальна техніка», «Кібернетика» в реферативній базі даних «Україніка наукова». Визначено основні параметри мережі та вузлів. Показано, що застосування характеристик центральності вузлів дає змогу визначити авторів наукових робіт, що вносять вагомий внесок у галузь, є керівниками наукових шкіл, підтримують зв'язок між науковими колективами.

Ключові слова: *мережі співавторів, реферативна база даних «Україніка наукова», формат УКРМАРК, наукова взаємодія, параметри вузлів мережі.*

Вступ

Оцінка наукової діяльності, в тому числі у співпраці є актуальною задачею в Україні та світі [1]. Багатоавторність у наукових публікаціях — один із найбільш частих об'єктів дослідження наукометрії, який відображає результати наукової співпраці.

Наукова комунікація є рушійною силою розвитку науки у сучасному світі. Обмін знаннями, досвідом, ідеями, ресурсами, інформацією, політичні та економічні чинники, підвищення якості та продуктивності результатів, індивідуальні фактори та просто задоволення від роботи спонукають науковців до співпраці [2, 3]. Постійне збільшення кількості спільних досліджень за кордоном було зафіксовано неодноразово [4].

Результати наукової співпраці частіше проявляються у написанні спільних наукових робіт, патентів на винаходи, виступах на конференціях. Тому аналіз сумісних наукових робіт надає найбільш об'єктивні та видимі результати співпраці. Комунікативність і впливовість окремих науковців найчастіше досліджують за допомогою побудови мереж співавторів і розрахунку параметрів вузлів співавторів.

Під науковою взаємодією також прийнято вважати становлення наукових шкіл. Наукова школа — неформальний творчий колектив дослідників різних поколінь, об'єднаних загальною програмою та стилем дослідницької роботи, які діють під керівництвом визнаного лідера [5]. Загальновизнано, що феномен «наукової школи» є основою для навчання науковців, розвитку наукових напрямів і базовою складовою української науки.

Метою статті є визначення основних характеристик і закономірностей мережі співавторства з комп'ютерних наук.

Характеристики складних мереж

Мережі співавторства є одним із прикладів застосувань концепції мережевої парадигми, що є ефективним інструментом дослідження складних систем [6].

Термін «складні мережі» виник на початку цього століття та включає мережі більш складної архітектури ніж класичні графи або випадкові мережі із заданою кількістю вузлів. Звичайно в таких мережах є невелика кількість вузлів з великою кількістю зв'язків — хабів (ядер), які значною мірою і визначають властивості даних мереж [6].

Для розрахунку характеристик мережі в цілому використовуються такі параметри як: кількість вузлів; кількість ребер; середня відстань між вузлами; діаметр мережі — найбільша геодезична відстань у мережі; щільність мережі — відношення кількості ребер у мережі до максимально можливої кількості ребер. Визначення клік (підгруп чи кластерів, в яких вузли зв'язані між собою сильніше, ніж членами інших клік), виділення компонент (частин мережі, що пов'язані всередині і не пов'язані між собою), знаходження перемичок (вузли, що при їхньому вилученні призводять до розпадання мережі) є одними з актуальних задач дослідження складних мереж [7].

Розділеність мережі на групи оцінюється коефіцієнтом кластеризації, що відображає відношення кількості зв'язків між сусідами до повної можливої кількості таких зв'язків. Загальний коефіцієнт кластеризації графа обчислюється як

$$C = \frac{1}{N}\sum_{i=1}^{N}\frac{E_i}{k_i(k_i-1)}, \qquad (1)$$

де N — кількість вершин; k_i — кількість зв'язків i-го вузла; E_i — кількість вузлів, сусідніх до i-го вузла, що зв'язані напряму. Чим ближче значення коефіцієнта до 1, тим більша вірогідність кластерної структури [8].

Однією з важливих характеристик графа є модулярність елементів і графа в цілому. Модулярність вузла — це величина, що оцінює щільність зв'язків у зв'язній компоненті порівняно зі зв'язками між компонентами. В загальному вигляді модулярність можна визначити як

$$Q = \sum_{i=1}^{N}\left(e_{ii} - a_i\right), \qquad (2)$$

де e_{ij} — елемент матриці суміжності графа, рівний відношенню кількості ребер, які поєднують два товариства i та j, до загальної кількості ребер у мережі; $a_i = \sum_{j=1}^{N} e_{ij}$ — відношення кількості ребер, що з'єднують вершини в товаристві i, до загальної кількості ребер. Висока модулярність мережі свідчить про сильний зв'язок у товариствах — кластерах і слабкий зв'язок самої мережі [9].

У теорії складних мереж існує декілька типів коефіцієнтів важливості вузлів, що визначаються як рівень їхньої центральності у графі. Причому деякі з концепцій були засновані на основі теорії складних мереж, інші ж вийшли із соціологічних досліджень. Існує чотири основних типи центральності, що широко використовуються в аналізі мереж: рівень центральності (degree centrality), центральність у розумінні посередництва (betweenness centrality), центральність у розумінні близькості (closeness centrality) та центральність власного вектора (eigenvector centrality).

Рівень центральності оцінює з якою кількістю інших учасників пов'язана конкретна особа, що для мереж співавторства також можна розуміти як степінь наукової взаємодії. У найпростішому випадку — це степінь конкретної вершини

$$C_D(i) = \sum_{j=1}^{N} m_{ij}, \tag{3}$$

де $m_{ij} = 1$, якщо вершина i пов'язана із вершиною j, та $m_{ij} = 0$ — в іншому випадку [10].

Степінь центральності характеризує авторів з точки зору комунікабельності та може використовуватися для передбачення продуктивності автора. За даними досліджень ця характеристика не корелює з середньою цитованістю та не повною мірою відображає всі аспекти комунікабельності авторів [10]. До недоліків даного індикатора для визначення комунікативності є відсутність врахування ваг ребер, тобто кількості сумісних публікацій авторів.

У роботі [11] запропоновано коефіцієнт для обрахунку центральності у зваженому графі для конкретної вершини:

$$C_D^{\omega\alpha}(i) = k_i^{(1-\alpha)} s_i^{\alpha}, \tag{4}$$

де враховується $k_i = \sum_{j=1}^{N} m_{ij}$ — сума зв'язків з іншими вершинами та $s_i = \sum_{j=1}^{N} \omega_{ij}$ — сума ваг відповідних зв'язків, α є коефіцієнтом, що підбирається залежно від конкретних випадків.

Центральність у розумінні близькості визначає наскільки близько вершина розташована відносно до інших. Якщо автор центральний за даною характеристикою, то він знаходиться у центрі досліджень певного напрямку та може осягнути його всебічно за допомогою своїх співавторів або через них швидше за інших отримати необхідні для цього знайомства

$$C_C(i) = \left[\sum_{j=1}^{N} d(i,j) \right]^{-1}, \; i \neq j, \tag{5}$$

де $d(i,j)$ — найменша відстань між вершинами *i* та *j* [12].

Центральність у розумінні посередництва визначає вершину, що зв'язує між собою підграфи. В розумінні наукової співпраці посередництво надає змогу визначити авторів, що утворюють зв'язок між науковими школами:

$$C_B(i) = \sum_{j<k} g_{jk}(i), \;\; i \neq j, \; k. \tag{6}$$

Тут $g_{jk}(i)$ — число найкоротших шляхів у графі, що проходять через *i*-ту вершину [12].

Метрика центральності власного вектора обчислюється за допомогою врахування ваг сусідніх вузлів, надає змогу оцінити вузли, що пов'язані з важливими сусідами, та обчислюється як нормалізоване значення власного вектора матриці суміжності. Інша метрика центральності — ексцентриситет — відображує можливість доступу до вузла від інших вузлів мережі, визначається як відстань від заданої вершини до найбільш віддаленої від неї. Алгоритм PageRank також дозволяє визначати, як часто, слідуючи зв'язками, можна попасти в один і той самий вузол, тобто найбільш комунікабельних авторів. PageRank і коефіцієнт авторитетності (Authority) були спроектовані для оцінки зв'язків між інтернет-сайтами. Коефіцієнт авторитетності визначає важливість вузла на основі ваг зв'язків, що, в свою чергу, ітеративно обчислюються на основі ваг вузлів [13].

Визначення важливих вузлів у мережі є актуальною задачею та потребує детального вивчення предмету дослідження, адже існує багато коефіцієнтів, що надають різнобічні характеритики вершин, та доцільність їхнього застосування визначається тільки відповідністю цілям експериментів.

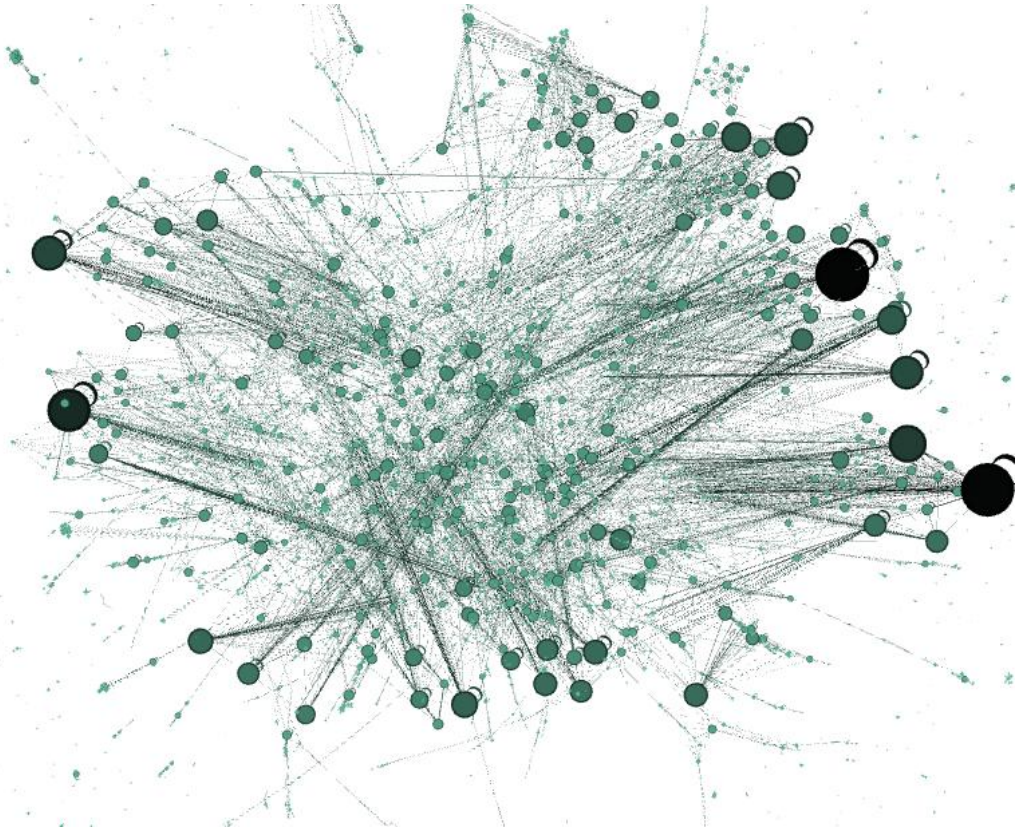
Мережі співавторства

Для побудови мереж співавторства у галузях «Інформаційна та обчислювальна техніка», «Кібернетика» в реферативній базі даних «Україніка наукова» був оброблений файл реферативної інформації за листопад 2012 року, що містив близько 440 000 записів. До рубрик «Інформаційна та обчислювальна техніка», «Кібернетика» входять теми: «Основи інформатики та обчислювальної техніки», «Аналогова і гібридна обчислювальна техніка», «Цифрова обчислювальна техніка», «Комп'ютери і програмування», «Кібернетичні моделі», «Теорія інформації», «Системний аналіз», «Теорія автоматів» та «Біоніка» [14]. Стандартний запис реферату наукової публікації формується на основі формату УКРМАРК, містить описову інформацію та власне розбитий на поля запис [15, 16]. Було створено програмні засоби мовою високого рівня Python, що виконували фільтрацію співавторів за індексами рубрикатора НБУВ «З97. Інформаційна та обчислювальна техніка», «З81. Кібернетика» та формували мережу співавторів. Візуалізація мережі була виконана за допомогою засобів програмного продукту Gephi [13].

Реферативна база даних «Україніка наукова» містить записи українською, англійською та російською мовами, тому інформація про публікації авторів може містити похибки. Для корекції результатів було взято до уваги лише перший ініціал автора, замінено у прізвищах та ініціалах українські літери «і», «ї», «є», на «и» та «е» відповідно, публікації деяких авторів трьома мовами були поєднані. Ваги зв'язків між авторами для кожної публікації визначалися обернено пропорційно до кількості її авторів.

У результаті отримано мережу співавторів, що містить 7740 авторів і 16287 зв'язків. Діаметр графа, тобто найбільша відстань між двома співавторами через інших авторів, рівний 22. Причому середня довжина шляху між вузлами близька до 8. Середня кількість людей, з якими співпрацює один співавтор, або середня степінь вузла, приблизно рівна 4. Середня зважена степінь близька до 3, тобто у співавторстві один автор має 3 публікації. На рисунку представлено мережу співавторів, де величини вузлів проранжовані за кількістю публікацій у співавторстві.

Мережа являє собою слабко зв'язані між собою підграфи (щільність графа = = 0,001). Модулярність мережі співавторів рівна 0,934, це свідчить про активну взаємодію в невеликих наукових групах, школах і слабкий зв'язок між ними (якщо модулярність > 1, то мережа містить меншу кількість більших за розміром клік). Наявність кластерної структури у досліджуваному графі підтверджується також середнім коефіцієнтом кластеризації, що рівний 0,823.



Мережа співавторства, величини вузлів проранжовані за кількістю публікацій у співавторстві

Усього в графі було 804 кліки та 672 слабко зв'язні компоненти. Наявність наукових шкіл у галузі «Комп'ютерних наук» можна простежити наочно за особливістю більшості клік, що в своїй основі містять потужного автора з великою кількістю статей у співавторстві та значну кількість маленьких вузлів — авторів-учнів (див. рис.). Для мереж співавторів також було виявлено авторів, що найчастіше публікують матеріали досліджень з одними і тими співавторами і зрідка можуть працювати з іншими. Характеристики центральності складних мереж не враховують дану особливість. Доцільно сформулювати показник центральності, що знижує значимість такого виду співпраці та враховує інші особливості наукової комунікації.

Список прізвищ, що мають максимальні показники за кількістю зв'язків і кількістю публікацій у співавторстві, представлено у табл. 1.

Таблиця 1. Автори наукових робіт з комп'ютерних наук, що мають найбільшу кількість співавторів і публікацій

Прізвище, ім'я	Кількість зв'язків	Прізвище, ім'я	Кількість статей (рівень центральності)
Харченко В.	81	Кожемяко В.	75
Кожемяко В.	80	Хорошко В.	70
Хорошко В.	65	Кузнецов А.	55
Мартынюк Т.	56	Бодянский Е.	49
Кузнецов А.	53	Харченко В.	48
Сергиенко И.	51	Сергиенко И.	47
Петров В.	50	Баркалов А.	47
Бодянский Е.	47	Глушаков С.	46
Крючин А.	45	Дорошенко А.	45
Палагин А.	45	Палагин А.	42
Кравец В.	43	Бидюк П.	42
Бидюк П.	41	Азаров О.	41

Ранжування за кількістю зв'язків і рівнем центральності відмінні, адже на початковій стадії обробки інформації з реферативної бази даних було визначено ваги зв'язків, що пропорційно розподілені між усіма авторами кожної публікації. Рівень центральності, що фактично є показником кількості статей у співпраці відображає об'єм напрацювань автора, а кількість зв'язків автора характеризують коло його співавторів.

Прізвища авторів також були проранжовані за характеристиками центральності, результати чого представлені у табл. 2.

Перелік прізвищ авторів у табл. 1 і 2 при врахуванні більшого інтервалу за найбільшими значеннями рівня центральності, центральності як посередництва, центральності за власним вектором і кількості зв'язків майже повторюється. Це підтверджує їхню значимість у наукових досягненнях і високий рівень співпраці в галузях «Інформаційна та обчислювальна техніка», «Кібернетика». Центральність як близькість виявляє авторів, що є близькими до основних спеціалістів галузі, і мають можливості та перспективи розвитку. Ранжування за значенням ексцентри-

ситету є близьким до останньої характеристики. Коефіцієнти PageRank та авторитетності ще раз підтвердили значимість початкових лідерів.

Таблиця 2. Автори, що є найбільш центральними в мережі співавторства

Прізвище, ім'я	Центральність як посередництво	Прізвище, ім'я	Центральність за власним вектором	Прізвище, ім'я	Центральність як близькість
Харченко В.	1341704	Кожемяко В.	1	Кудрявцева Н.	15.2079
Бодянский Е.	1130606	Харченко В.	0.980745	Филь И.	15.2079
Кузнецов А.	1054234	Петров В.	0.815397	Меркулова Е.	14.21019
Палагин А.	877582.8	Крючин А.	0.736181	Коков А.	14.21019
Анисимов А.	731121.4	Мартынюк Т.	0.653736	Голиков В.	14.20867
Гриценко В.	668846.4	Кравец В.	0.633307	Тихонов В.	14.20829
Мельник А.	645622.2	Палагин А.	0.588672	Ли И.	14.09603
Хорошко В.	630350.4	Браткевич В.	0.535928	Карпов А.	14.09603
Бидюк П.	625496.8	Плоткин В.	0.526162	Трунов В.	14.09603
Кожемяко В.	624592.2	Кузнецов А.	0.52596	Шилова Е.	14.09603
Сергиенко И.	590181.9	Костюкевич С.	0.51239	Козловский Е.	13.26155
Жуков И.	572185.7	Климнюк В.	0.504692	Лякутин В.	13.26155

Висновки

У роботі досліджено мережу співавторства в галузях «Інформаційна та обчислювальна техніка» та «Кібернетика» в реферативній базі даних «Україніка наукова». Визначено основні параметри вузлів і мережі, такі як діаметр графа, загальне та середнє значення кількості вузлів і зв'язків, модулярність, коефіцієнт кластеризації, показники центральності вузлів та ін.

Показано, що побудована мережа співавторів являє собою слабко зв'язані між собою підграфи. Автори наукових публікацій частіше взаємодіють у невеликих наукових групах, школах, що слабко пов'язані між собою. Наявність наукових шкіл у досліджуваній галузі можна простежити наочно за особливістю більшості підгруп, що в своїй основі містять потужного автора з великою кількістю статей у співавторстві та значну кількість маленьких вузлів — учнів.

Показано, що застосування характеристик центральності вузлів дає змогу визначити авторів наукових робіт, що вносять вагомий внесок у галузь, є керівниками наукових шкіл, підтримують зв'язок між науковими колективами.

Запропоновано створити показник центральності, що враховує особливості наукової співпраці в галузі «Комп'ютерні науки» і створювати умови та більш активно взаємодіяти між організаціями для перерозподілу знань і досвіду. Доцільним також є поширення практики наукової співпраці на міждержавному рівні. Підвищення рівня конкурентоспроможності вітчизняної науки можливо за умови комунікації науковців із зарубіжними науковими групами та участь у сумісних проектах, які на даний момент є найбільш актуальними у світовій практиці.

Галузі «Інформаційна та обчислювальна техніка» та «Кібернетика» потребують подальших наукометричних досліджень і визначення наукових шкіл, перспективних наукових напрямів досліджень для забезпечення її стабільного розвитку.

1. *Копанєва Є.О.* Національні індекси наукового цитування / Є.О. Копанєва // Бібліотечний вісник. — 2012. — № 4(210). — С. 31–33.

2. *Bukvova H.* Studying Research Collaboration: A Literature Review [Електронний ресурс] / Helena Bukvova // Sprouts: Working Papers on Information Systems. — 2010. — N 10(3). — P. 1–19. — Режим доступу: http://sprouts.aisnet.org/10-3. — Назва з екрану.

3. *Шубина Н.Л.* Научная коммуникация: поиски разумного компромисса // Известия РГПУ им. А.И. Герцена. — 2009. — № 104. — С. 87–97.

4. *Examining* Core Elements of International Research Collaboration: Summary of a Workshop. — National Research Council. — Washington, DC: The National Academies Press, 2011. — 128 p.

5. *Довбня П.* Наукові школи та їх класифікація / Петро Довбня, Ірина Доброскок // Гуманітарний вісник ДВНЗ «Переяслав-Хмельницький державний педагогічний університет ім. Г. Сковороди». — 2008. — № 16. — С. 57–60.

6. *Евин И.А.* Введение в теорию сложных сетей / И.А. Евин // Компьютерные исследования и моделирование. — 2010. — Т. 2, № 2. — С. 121–141.

7. *Ландэ Д.В.* Интернетика: Навигация в сложных сетях: модели и алгоритмы / Д.В. Ландэ, А.А. Снарский, И.В. Безсуднов. — М.: Либроком (Editorial URSS), 2009. — 264 с.

8. *Latapy M.* Main-Memory Triangle Computations for Very Large (Sparse (Power-Law)) / Matthieu Latapy // Graphs, in Theoretical Computer Science (TCS). — 2008. — Vol. 407, N 1–3. — P. 458–473.

9. *Апанович З.В.* Средства визуального анализа информационного наполнения порталов, входящих в облако Linked Open Data / З.В. Апанович, П.С. Винокуров, Т.А. Кислицина // Труды 13-й Всеросийской науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL'2011. — Воронеж, Россия, 19–22 октября 2011 г. — CEUR, 2011. — Т. 803. — С. 113–120.

10. *Liao Ch. Hs.* Quantifying the Degree of Research Collaboration: A Comparative Study of Collaborative Measures / Chien Hsiang Liao, Hsiuju Rebecca Yen // Journal of Informetrics. — 2012. — Vol. 6, N 1. — P. 27–33.

11. *Opsahl T.* Node Centrality in Weighted Networks: Generalizing Degree and Shortest Paths / Tore Opsahl, Filip Agneessens, John Skvoretz // Social Networks. — 2010. — Vol. 32, N 3. — P. 245–251.

12. *Alireza A.* Betweenness Centrality as a Driver of Preferential Attachment in the Evolution of Research Collaboration Networks / Abbasi Alireza, Hossain Liaquat, Loet Leydesdorff // Journal of Informetrics. — 2012. — Vol. 6, N 3. — P. 403–412.

13. *An* Open-Source Software for Visualizing and Analyzing Large Network Graphs «Gephi» [Електронний ресурс]. — Режим доступу: http://gephi.org. — Назва з екрану.

14. *Реферативна* база даних «Україніка наукова» [Електронний ресурс]. — Режим доступу: http://nbuv.gov.ua/db/ref.html. — Назва з екрану.

15. *Ланде Д.В.* Наукометричні дослідження мереж співавторства по базі даних «Україніка наукова» / Д.В. Ланде, І.В. Балагура // Реєстрація, зберігання і оброб. даних. — 2012. — Т. 14, № 4. — С. 41–51.

16. *Антоненко І.П.* Каталогізація електронних ресурсів: наук.-метод. посіб. / І.П. Антоненко, О.В. Баркова; Нац. б-ка України ім. В.І. Вернадського. — К.: НБУВ, 2007. — 115 с.

Надійшла до редакції 05.03.2013