УДК 501

ТОПОЛОГИЯ ГРАФА СОАВТОРСТВА В ОБЛАСТИ ФИЗИКИ В РОССИИ

О. В. Иванов[1], А. М. Коваленко[1], А. В. Колобов[1], В. В. Королева[1], А. В. Леонидов[1,2], Е. Е. Серебрянникова[1,2]

В представленной работе рассматриваются топологические свойства графа соавторства применительно к исследованиям по физическим наукам в России за 2012–2018 *годы. Предметом изучения являются два различных взвешенных графа – в первом веса ребер соответствуют количеству совместных статей двух исследователей, а во втором веса рассчитываются по методике, учитывающей качество журналов, в которых опубликованы совместные статьи. На основе анализа вектора центральности PageRank показано, что ранжирование авторов в данных двух графах значительно отличается. Однако в обоих графах лидерами ежегодно становятся преимущественно регулярно публикующиеся исследователи.*

Ключевые слова: граф соавторства, физические науки, топология, PageRank.

В представленной работе рассматриваются топологические свойства графа соавторства применительно к исследованиям по физическим наукам в России за 2012–2018 годы, построенного на основе информации о публикационной активности российских исследователей из базы данных WebOfScience. Исследования графа соавторства являются одной из важных точек приложения теории сложных сетей к количественному анализу различных явлений социально-экономической реальности, см. напр. [1]. Начало масштабным исследованиям графа соавторства с привлечением объемных баз данных было положено в работе [2] и с тех пор активно продолжается [3]. Свойства графа соавторства отражают особенности одного из важнейших составных элементов процесса генерации научного знания – объединения усилий индивидуальных исследователей

[1] ФИАН, 119991 Россия, Москва, Ленинский пр-т, 53; e-mail: leonidovav@lebedev.ru.

[2] МФТИ, 141701 Россия, Московская область, г. Долгопрудный, Институтский пер., 9.

для осуществления совместной работы, результатом которой является публикация в реферируемом научном издании.

Определим граф соавторства $\mathcal{G}_t$ как граф, вершинами которого являются индивидуальные исследователи[1], а ребра между вершинами возникают при наличии совместных публикаций у соответствующих пар исследователей за некоторый рассматриваемый период времени, например, год, с индексом t. Существенный интерес представляет также анализ модификаций исходного графа $\mathcal{G}$, учитывающих интенсивность и/или качество сотрудничества авторов совместных публикаций, отвечающих соответствующим определенным ниже взвешенным графам.

Ниже будут использованы следующие обозначения:

1. t – год, за который рассматривается граф (в данных $t = 2012, \ldots, 2018$);

2. R_t – количество вершин графа $\mathcal{G}_t$, т.е. количество исследователей, являющихся соавторами хотя бы одной статьи за год t;

3. $\mathcal{R}_t = \{\mathrm{id}_1^t, \ldots, \mathrm{id}_{R_t}^t\}$ – множество вершин графа $\mathcal{G}_t$;

4. $\mathcal{G}_t$ – невзвешенный граф соавторства, ребро в котором демонстрирует наличие хотя бы одной совместной публикации у исследователей в году t;

5. $\mathcal{G}_t = (\mathcal{R}_t, \mathcal{E}_t)$, где $\mathcal{E}_t$ – множество ребер графа $\mathcal{G}_t$;

6. матрица $\mathbf{G}_t = \{g_{ij}^t\}, i, j = 1, \ldots, R_t$, – матрица смежности графа соавторства $\mathcal{G}_t$, в ней $g_{ij}^t = 1$, если исследователи id_i^t и id_j^t имели хотя бы одну общую публикацию в году t, и $g_{ij}^t = 0$, если это не так;

7. $\mathcal{N}_t$ – взешенный граф соавторства, в котором вес ребра равен количеству совместных статей соединяемых им узлов в году t;

8. матрица $\mathbf{N}_t = \{n_{ij}^t\}, i, j = 1, \ldots, R_t$, – матрица весов графа $\mathcal{N}_t$, для которой n_{ij}^t равно количеству совместных публикаций в году t у исследователей id_i^t и id_j^t;

9. $\mathcal{W}_t$ – взешенный граф соавторства, в котором вес ребра равен сумме фракционных баллов совместных статей в году t;

10. матрица $\mathbf{W}_t = \{w_{ij}^t\}, i, j = 1, \ldots, R_t$, – матрица весов графа $\mathcal{W}_t$, для которой w_{ij}^t рассчитывается по следующей формуле:

$$w_{ij}^t = \sum_{k=1}^{n_{ij}^t} \frac{2Q_k}{\nu_k}, \tag{1}$$

[1]В граф входят только те исследователи, у которых были работы в соавторстве с другими исследователями. Если исследователь публиковал статьи только в одиночку, то он не появится в графе и эти статьи никак не будут учтены.

где Q_k – это вес статьи k (зависящий от качества журнала, в котором опубликована данная статья, см. табл. 1), а ν_k – это количество авторов у статьи k. Отметим, что взвешенный граф с весами ребер вида (1) ранее не изучался.

Таблица 1

Таблица весов статей в формуле (1).
В таблице использованы следующие обозначения для изданий:
Q1–Q4 – индексируемые Web of Science (WoS), Q — без квартиля из Web of Science Core Collection (WoSCC), S – индексируемые в Scopus, но не в WoS и WoSCC, R – индексируемые в RSCI WoS, но не в WoSCC, V – входящие в список BAK, B – зарегистрированные в Российской книжной палате

Q1	Q2	Q3	Q4	Q	S	R	V	B
19.7	7.3	2.7	1	1	1	0.75	0.5	1

Начнем наш анализ графа соавторства $\mathcal{G}_t$ с самых общих характеристик. Динамика количества вершин графа, количества его ребер и плотности

$$\rho_t = 2|\mathcal{E}_t|/R_t(R_t-1),$$

т. е. доли имеющихся ребер по сравнению с полным графом с тем же числом вершин (здесь и далее под $|\cdot|$ понимается мощность соответствующего множества), представлена в табл. 2. Из данных величин следует, что при существенном росте числа вершин и ребер графа его плотность упала за рассматриваемый период в 2.7 раза.

Таблица 2

Эволюция количества вершин R_t, количества ребер $|\mathcal{E}_t|$, плотности ρ_t, а также абсолютного и относительного размера LCC графа $\mathcal{G}_t$ за 2012–2018 *гг.*

год	R_t	$\lvert\mathcal{E}_t\rvert, x10^{-4}$	Плотность (ρ_t)	$\lvert LCC\rvert$	$\lvert LCC\rvert/R_t$
2012	18638	130729	7.5	9208	49%
2013	19097	112717	6.2	9292	49%
2014	21711	120967	5.1	11961	55%
2015	27461	171515	4.5	16738	61%
2016	32013	204563	4.0	20246	63%
2017	36135	213651	3.3	21009	58%
2018	34390	202475	3.4	20354	59%

Важнейшей характеристикой графа соавторства является размер наибольшего связного кластера (LCC, Large Connected Cluster). Данные по эволюции размера LCC также приведены в табл. 2. Из данных результатов мы видим, что за рассматриваемый период рос не только абсолютный, но и, что еще более важно, относительный размер LCC, который за рассматриваемый период вырос в 1.2 раза. Тем самым, объем научного сотрудничества, приводящий к подготовке совместных публикаций, за рассматриваемый период существенно вырос. Стоит отметить, что размер второго по величине связного кластера во всех периодах составляет доли процента от размера LCC.

Отметим также, что общими для всех семи графов $\{\mathcal{G}_t,\ t = 2012, \ldots, 2018\}$ являются 3469 вершин, т. е. на протяжении семи лет только 3469 авторов ежегодно публиковали статью в соавторстве. При этом, если учитывать только статьи с не более чем 20-ю авторами, то таких "регулярно печатающихся" авторов будет 2678. На рис. 1 приведена диаграмма, демонстрирующая распределение числа авторов по количеству лет, в которых у них имеются публикации в соавторстве за период 2012–2018 гг. Из данной диаграммы следует, что большая часть авторов (56%) имеют публикации лишь в одном из семи рассматриваемых годовых периоде.

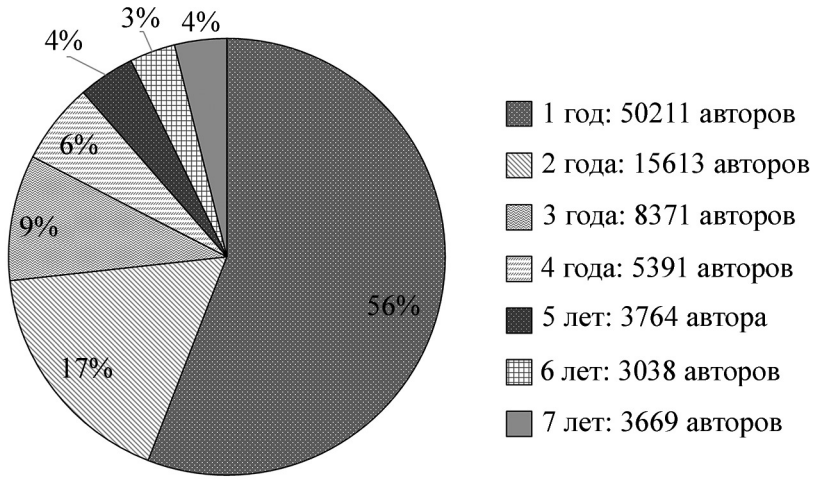

Рис. 1: *Доли авторов, имеющих статьи в соавторстве в n годах за период с* 2012 *по* 2018 *гг., где* $n = 1, \ldots, 7$. *На рис.* 1 *также справочно приведены соответствующие количества авторов.*

Ранжирование узлов по их важности с точки зрения роли в совместном написании статей и обеспечении их качества можно осуществить с использованием вектора

центральности[2] PageRank [4] $\mathbf{x} = \{x_i\}$, который определен уравнением

$$x_i = d\sum_j \frac{Z_{ij}}{\sum_k Z_{kj}} x_j + \beta = d\sum_j \gamma_{ij} x_j + \beta, \tag{2}$$

или эквивалентно

$$\mathbf{x} = \beta(I - d\Gamma)^{-1}\mathbf{1}, \tag{3}$$

где I – это единичная матрица, $\mathbf{1}$ – вектор, состоящий из единиц, $\Gamma = \{\gamma_{ij}\}$, $i, j = 1, \ldots, N$, – это стохастическая слева матрица[3], построенная на основе исходной матрицы весов Z графа (в рассматриваемом случае матрицы W_t и N_t). Уравнение (3) определяет центральность PageRank с параметрами β и d. Отметим, что из соотношения (3) следует, что значение параметра β не влияет на ранжирование, данный параметр является нормировочным. При вычислениях параметр d был выбран, как и в работе [4], равным 0.85.

Существенный интерес представляет сравнительный анализ результатов ранжирования по фракционному счету и числу совместных статей. Удобный способ такого сравнения дает вычисление коэффициента ранговой корреляции Кендалла[4] c_K, результаты которого приведены в табл. 3.

Таблица 3

Величина коэффициента ранговой корреляции Кендалла c_K векторов PageRank графов $\mathcal{N}_t$ и $\mathcal{W}_t$, $t = 2012, \ldots, 2018$.

Год	2012	2013	2014	2015	2016	2017	2018
c_K	0.41	0.42	0.38	0.37	0.28	0.25	0.28

Мы видим, что отличие в ранжировании для графов $\mathcal{N}_t$ и $\mathcal{W}_t$ за рассматриваемый период существенно выросло.

Графами на регулярных вершинах будем называть графы $\tilde{\mathcal{N}}_t$ и $\tilde{\mathcal{W}}_t$, $t = 2012, \ldots, 2018$, являющиеся подграфами исходных графов $\mathcal{N}_t$ и $\mathcal{W}_t$ на 3469-ти вершинах, отвечающих регулярно печатающимся авторам.

[2]Центральностью графа называют некоторый способ ранжирования узлов по тому или иному критерию важности.

[3]Стохастической слева называют матрицу, сумма элементов каждого столбца которой равна единице.

[4]Коэффициент ранговой корреляции Кендалла для двух векторов ранжирования отражает степень схожести соответствующих иерархий их компонент.

Если рассматривать только лидеров по PageRank, например TOP-20, то оказывается, что порядка половины лидеров являются одновременно лидерами и в графе $\mathcal{N}_t$, и в графе $\mathcal{W}_t,\ t = 2012, \ldots, 2018$ (см. табл. 4). При этом 90% (31 из 34) авторов, являющихся лидерами в обоих графах, оказываются регулярно публикующимися, то есть относящимися к графам на регулярных вершинах.

Таблица 4

Мощность пересечения TOP-20 векторов PageRank графов $\mathcal{N}_t$ и $\mathcal{W}_t,\ t = 2012, \ldots, 2018$ (графы на всех вершинах) и графов $\tilde{\mathcal{N}}_t$ и $\tilde{\mathcal{W}}_t,\ t = 2012, \ldots, 2018$ (графы на регулярных вершинах)

	Год						
	2012	2013	2014	2015	2016	2017	2018
Графы на всех вершинах	9	9	11	8	7	8	11
Графы на регулярных вершинах	9	8	9	7	7	8	11

Интересно, что если рассматривать TOP-10 лидеров по PageRank, то оказывается, что почти все авторы, входящие в TOP-10 в течение рассматриваемых 7 лет являются регулярно печатающимися. Для графа $\mathcal{W}_t,\ t = 2012, \ldots, 2018$ 38 из 42 авторов, входивших хотя бы раз в TOP-10, являются регулярно печатающимися, а для для графа $\mathcal{N}_t,\ t = 2012, \ldots, 2018$ – 36 из 40.

Существенный интерес представляет анализ структуры списков TOP-10 с точки специализации авторов по областям физики. С этой целью каждому автору в списке TOP-10 за рассматриваемый год сопоставляется соответствующий список статей, по которому формируется список журналов (тем самым, журнал может входить в такой список несколько раз). Каждому журналу сопоставляется список относящихся к нему областей физики. Объединение таких списков за 2012–2018 гг. и стало предметом нашего изучения. Как уже упоминалось ранее, в работе изучались два способа ранжирования авторов – по числу статей и по фракционному баллу. На рис. 2 приведено распределение по количеству упоминаний в этом списке 20 наиболее упоминаемых областей физики, отвечающих обоим упомянутым способам ранжирования. Помимо ожидаемого лидерства междисциплинарной тематики, отметим лидирующие позиции физики твердого тела, оптики и физики нанотехнологий.

В заключение перечислим основные результаты, описанные в настоящей работе. При рассмотрении графа соавторства работ в области физики в России за 2012–2018 гг. нами были изучены следующие его характеристики:

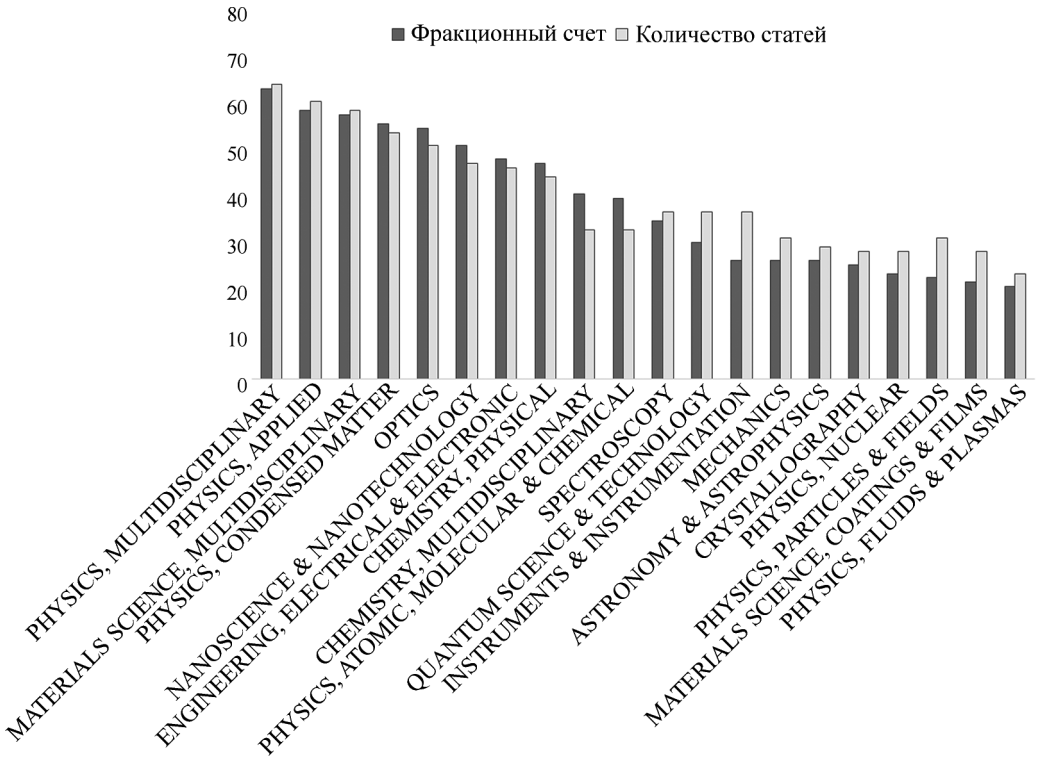

Рис. 2: *ТОП*-20 *наиболее часто упоминаемых областей физики, отвечающих работам TOP*-10 *лидеров по PageRank за* 2012–2018 *гг.*

– размер наибольшего связного кластера, который за рассматриваемый период существенно вырос;

– ранжирование узлов по PageRank и изучение устойчивости такого ранжирования по годам, продемонстрировавшего сильные ежегодные изменения;

– степень представленности в публикациях, отвечающих спискам TOP-10 по PageRank, различных областей физики.

Работа поддержана грантом Министерства науки и высшего образования № 05.601.21.0020 (уникальный идентификатор соглашения RFMEFI60119X0020) “Исследование механизмов адаптивного формирования кадрового потенциала для проведения разномасштабных программ исследований по приоритетным направлениям научно-технологического развития Российской Федерации”.

Л И Т Е Р А Т У Р А

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

[2] M. E. J. Newman, *Coauthorship networks and patterns of scientific collaboration.* Proceedings of the national academy of sciences **101**(1), 5200 (2004). DOI: 10.1073/pnas.0307545100.

[3] S. Kumar, Aslib Journal of Information Management **67**(1), 55 (2015). DOI:10.1108/AJIM-09-2014-0116.

[4] S. Brin and L. Page, Computer Networks and ISDN Systems **30**(1-7), 107 (1998). DOI: 10.1016/S0169-7552(98)00110-X.

Поступила в редакцию 15 июня 2020 г.
После доработки 20 июня 2020 г.
Принята к публикации 21 июня 2020 г.