

THE HYPER-NETWORK OF SCIENTIFIC CO-AUTHORSHIP. DB REPEC DATA ANALYSIS

S. V. Bredikhin, V. M. Lyapunov, N. G. Scherbakova

Institute of Computational Mathematics and Mathematical Geophysics SB RAS,
630090, Novosibirsk, Russia

DOI: 10.24412/2073-0667-2022-4-70-83

EDN: MJFKPC

The article deals with the modeling of a complex network of co-authorship, presented in the form of a hypergraph. The formal information necessary to describe multiple relationships between coauthors is given and two models of the analyzed object are presented. Based on real information extracted from the bibliographic database RePEc, a hypergraph of the co-authorship network is constructed.

Most of the previous studies consider the co-authorship relation between two authors as a collaboration. So a network is represented as a simple graph in which link relates only a pair of authors that are coauthors of at least one scientific paper (SP). These pairwise networks have been studied from many aspects such as degree distribution analysis, community extraction, authors ranking, see, for example, [4–8]. Such networks does not provide a complete description of the collaboration because we only know whether scientists have collaborated or not but we can't know whether a group of authors linked together in the network were coauthors of the same paper or not. As a variant of the representation that takes into account n -ary relations between authors, a bipartite graph may be considered, in which one partite set represents the authors, the other — SPs prepared by these authors. This makes it possible to use the apparatus of graph theory, but at the same time, heterogeneity in the definition of nodes makes more complicated the study of such topological properties as connectivity and clustering. Therefore, in [10], it is proposed to use a graph generalization, a hypergraph [11], to represent a complex system and call it a hyper-network. Edges of a hypergraph can relate groups of more than two nodes.

A (undirected) hypergraph $H = (V, E)$ on a finite set $V = v_1, v_2, \dots, v_n$ is defined by the family $E = (E_1, E_2, \dots, E_m)$ of subsets of the set V . An element $v_i \in V$ is called a node, an element $E_i \in E$ is called an (hyper)edge [17]. Let $P = \{p_1, p_2, \dots, p_m\}$ be the set of SPs, and $S = \{s_1, s_2, \dots, s_n\}$ be the set of their authors. We assume that P contains only those SPs that have two or more authors, i.e. the constructed hypergraph will not have edges consisting of a single vertex. Let us define a hypergraph $H_1 = (V, E_1)$ such that the set S is mapped to the set of vertices V , and the set P is mapped to the set of edges E_1 , and if the SP p_i is prepared precisely by the authors v_1, v_2, \dots, v_k , then $E_i = \{v_1, v_2, \dots, v_k\}$ is an edge, $E_i \in E_1$. The number of edges $m_1 = |E_1|$ is the number of publications $|P|$ [10]. We can also define a hypergraph $H_2 = (V, E_2, w)$ in which nodes represent authors and hyper-edges represent the groups of authors that have published papers together. Here $E_i = \{v_1, v_2, \dots, v_k\} \in E_2$ if there is at least one SP jointly published by the authors v_1, v_2, \dots, v_k . The edge weight is the number of SPs published jointly by these k authors. Number of edges $m_2 = |E_2|$ is the number of groups of authors [6].

In our work, we consider a set of SPs indexed in the RePEc database at the time of extraction. The procedure for filtering “raw” data is presented in [15]. As a result, having 91113 co-authored SPs

This work was carried out under state contract with ICMMG SB RAS (0251-2021-0005).

and 32434 authors we construct the hypergraph $H^{ca} = (V, E)$ by analogy with H_1 above. At this stage we use the bipartite incidence graph $K(H^{ca}) = (V, V', E_K)$ in order to calculate a number of parameters of H^{ca} . The graph $K(H^{ca})$ that is isomorphic to H^{ca} can be obtained by associating with each hyper-edge $E_j \in E$ an additional vertex v_{e_j} and defining the set $V' = \{v_{e_j} : E_j \in E\}$ such that an edge between $v \in V$ and $v_{e_j} \in V'$ exists iff $v \in E_j$ [24]. It is shown that the hypergraph H^{ca} is neither simple nor conformal. Parameter values are given in Tab. 2. As an example, we consider the hypergraph component consisting of 12 nodes and 27 edges (Fig. 1, Tab. 1). It is noted that based on the hypergraph, co-authorship networks considered in the works [15, 16] can be built, the reverse is not true.

Key words: complex networks, scientific co-authorship, hypergraph, bipartite graph, bibliometry.

References

1. BOCCALETTI S., LATORA V., MORENO Y., CHAVEZ M., HWANG D. U. Complex networks: Structure and dynamics // *Phys. Rep.* 2006. V. 424, iss. 4–5. P. 175–308. DOI: 10.1016/j.physrep.2005.10.009.
2. BATTISTON F., CENCETTI G., IACOPINI I., LATORA V., LUCAS M., PATANIA A. YOUNG J.-G., PETRI G. Networks beyond pairwise interactions: Structure and dynamics // *Phys. Rep.* 2020. V. 874. P. 1–92. DOI: 10.1016/j.physrep.2020.05.004.
3. SHCHERBAKOVA N. G. Modelirovanie gruppovykh vzaimodejstvij kompleksnykh sistem. Obzor // *Problemy informatiki.* 2022. N. 3. S. 24–45.
4. NEWMAN M. E. J. Scientific collaboration networks. I. Network construction and fundamental results // *Phys Rev. E*, 64(1), 016131. DOI: 10.1103/PhysRevE.64.016131.
5. NEWMAN M. E. J. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality // *Phys. Rev. E*, 64(1), 016132. DOI: 10.1103/PhysRevE.64.016132.
6. SAVIČ M., IVANOVIĆ M., RADOVANOVIĆ M., OGNJANOVIĆ Z., PEJOVIĆ A. Exploratory analysis of communities in co-authorship networks: A case study // *Intern. Conf. on ICT Innovations*, Springer, 2019. P. 55–64.
7. BARABASI A. L., JEONG H., NEDA Z., RAVASZ E., SCHUBERT A., VICSEK T. Evolution of the social network of scientific collaborations // *Physica A.* 2002. V. 311. P. 590–614. DOI: 10.48550/arXiv.cond-mat/0104162.
8. UDDIN S., HOSSAIN L., ABBASI A., RASMUSSEN K. Trend and efficiency analysis of co-authorship network // *Scientometrics.* 2012. V. 90, N 2. P. 687–699. DOI: 10.1007/s11192-011-0511-x.
9. NEWMAN M. E. J., STROGATZ S. H., WATTS D. J. Random graphs with arbitrary degree distributions and their applications // *Phys. Rev. E* 64, 026118. 2001. DOI: 10.1103/PhysRevE.64.026118.
10. ESTRADA E., RODRIGUEZ-VELAZQUEZ J. A. Complex networks as hypergraphs // *Arxiv: physics/0505137*, 2005. DOI: 10.1016/j.physa.2005.12.002.
11. BERGE C. *Hypergraphs*. Amsterdam; N. Y.; Oxford; Tokyo: North-Holland, 1989.
12. TORRES L., BLEVINS A. S., BASSET D., ELIASSI-RAD T. The why, how, and when of representations for complex systems // *SIAM Rev.* 2021. V. 63, N 3. P. 415–485. DOI: 10.1137/20M1355896.
13. OUVREARD X., LE GOFF X.-M., MARCHAND-MAILLET S. Networks of collaborations: Hypergraph modeling and visualization // *ArXiv: 1809.00164v1*. DOI: 10.48550/arXiv.1809.00164.
14. HAN Y., ZHOU B., PEI J., JIA Y. Understanding importance of collaborations in coauthorship networks: A supportiveness analysis approach // *Proc. 2009 SIAM Intern. Conf. on Data Mining.* 2009. P. 1112–1123. DOI: 10.1137/1.9781611972795.95.

15. BREDIKHIN S. V., LYAPUNOV V. M., SHCHERBAKOVA N. G. Struktura i parametry nezvzveshennoj seti soavtorstva na osnove dannykh BD RePEc // Problemy informatiki. 2021. N. 3. S. 56–67. DOI: 10.24411/2073-0667-2021-3-56-57.
16. BREDIKHIN S. V., LYAPUNOV V. M., SHCHERBAKOVA N. G. Ranzhирование uzlov vzveshennoj seti soavtorstva: analiz dannykh BD RePEc // Problemy informatiki. 2021. N. 4. S. 5–15. DOI: 10.24412/2073-0667-2021-4-67-83.
17. VOLOSHIN V. I. Introduction to graph and hypergraph theory. N. Y.: Nova Science Publishers, Inc., 2009.
18. BRETTO A. Hypergraph theory: An introduction. Heidelberg: Springer Intern. Publishing, 2013. DOI: 10.1007/978-3-319-00080-0.
19. MARTINEZ M. G., STARK H. M., TERRAS A. A. Some Ramanujan hypergraphs associated to $GL(n, \mathbb{F}_q)$ // Proc. Am. Math. Soc. 2001. V. 129, P. 1623–1629. S. 0002-9939(00)05965-7.
20. OUVRARD X. Hypergraphs: An introduction and review // Arxiv: 2002.05014v2, 2020. DOI: 10.48550/arXiv.2002.05014.
21. ZHOU D., HUANG J., SCHÖKOPF B. Learning with hypergraphs: Clustering, classification, and embedding // Proc. 19th Internat. Conf. on Neural Inform. Proc. Syst. 2007. P. 1601–1608. DOI: 10.7551/mitpress/7503.003.0205.
22. BAHMANIAN M. A., SAJNA M. Connection and separation in hypergraphs // Theory and Appl. of Graphs. 2015. V. 2, iss. 2. DOI:10.20429/tag.2015.020205.
23. DISTEL' R. Teoriya grafov. Per. s angl. Novosibirsk: Izdatel'stvo instituta matematiki. 2002. 336 s. ISBN 5-86134-101-X.
24. ZYKOV A. A. Gipergrafy // Uspekhi mat. nauk. 1974. T. 29, vyp. 6. S. 89–156.
25. BORGATTI S. P., EVERETT M. G. Network analysis of 2-mode data // Social networks. 1997. V. 19. P. 243–269. DOI: 10.1016/S0378-8733(96)00301-2.
26. COOPER J., DUTLE A. Spectra of uniform hypergraphs // Lin. Algebra and Its Appl. 2012. V. 436. P. 3268–3292. DOI: 10.48550/arXiv.1106.4856.
27. BANERJEE A., CHAR A., MONDAL B. Spectra of general hypergraphs // Lin. Algebra and Its Appl. 2017. V. 518. P. 14–30. DOI: 10.1016/j.laa.2016.12.022.
28. KUMAR T., VAIDYANATHAN S., ANANTHAPADMANABHAN H. Hypergraph clustering: A modularity maximization approach // ArXiv: 1812.10869[cs.G]. DOI: 10.48550/arXiv.1812.10869.
29. KAMIŃSKI B., POULIN V., PRALAT P., SZUFEL P., THÉBERGE F. Clustering via hypergraph modularity // PLoS ONE. 2019. V. 14(11), e0224307. DOI: 10.1371/journal.pone.0224307.
30. ZHOU V., NAKHLEH L. Properties of metabolic graphs: biological organization or representation artifacts? // BMC Bioinformatics. 2011. V. 12, 132. DOI: 10.1186/1471-2105-12-132.

ГИПЕРСЕТЬ НАУЧНОГО СОАВТОРСТВА. АНАЛИЗ ДАННЫХ БД RERES

С. В. Бредихин, В. М. Ляпунов, Н. Г. Щербакова

Институт вычислительной математики и математической геофизики СО РАН,
630090, Новосибирск, Россия

УДК 519.177

DOI: 10.24412/2073-0667-2022-4-70-83

EDN: MJFKPC

Рассмотрены вопросы моделирования комплексной сети научного соавторства, представленной в виде гиперграфа, в отличие от традиционного подхода к изучению этого феномена, базирующегося на построении взвешенного либо невзвешенного графа. Приведены формальные сведения, необходимые для описания множественных отношений между группами соавторов, представлены две модели анализируемого объекта. На основе реальной информации, извлеченной из библиографической базы данных, сконструирован гиперграф сети соавторства, измерены его параметры и сформулированы основные свойства. Приведен содержательный пример. В результате работы феномен научного соавторства рассмотрен с новой точки зрения.

Ключевые слова: комплексные сети, научное соавторство, гиперграф, двудольный граф, библиометрия.

Введение. Неформальный термин «комплексные системы» объединяет естественные и искусственные объекты, имеющие сложные внутренние связи, влияющие на функционирование системы в целом. Интерес к их изучению возник в конце XX в. в результате анализа и моделирования реальных весьма непостоянных биологических, социальных и технических конструкций, таких как: погода, распределенные системы передачи информации, интернет вещей и т. п. Одним из методов исследования подобных объектов является представление в виде комплексной сети (КС), т. е. множества дискретных элементов и отношений между ними. Для моделирования КС принято использовать простые, ориентированные либо взвешенные графы, состоящие из множества вершин, соответствующих изучаемым элементам, часть которых объединена в пары на основе некоторого отношения. Методы исследования объектов с бинарными отношениями достаточно хорошо изучены (см., например, обзор [1]). Поскольку в реальных системах задействованы сложные многосторонние отношения между подсистемами, моделирование всех отношений как бинарных приводит к потере информации, присутствующей в исходном объекте. Поэтому для моделирования многосторонних зависимостей используются математические структуры, явно отражающие такие отношения (см., например, обзоры [2, 3]).

Соавторством называют форму объединения ученых с целью проведения совместных исследований, результатом которых обычно является общая публикация. Соавторами являются все авторы научной публикации (НП), внесшие вклад в ее подготовку и разделяющие ответственность за полученные результаты. Информация о соавторстве извлекается

Исследования выполнены в рамках государственного задания ИВМиМГ СО РАН (0251-2021-0005).

из библиографических баз данных (БД) и, как правило, содержит название НП, имена и аффилиации авторов, год издания. Эти сведения позволяют построить сеть соавторства N^{ca} , которую можно изучать как социальную сеть. Традиционный подход к исследованию N^{ca} , изложенный в основополагающих работах [4, 5], редполагает, что узлы соответствуют авторам, а ребра — отношению соавторства, устанавливаемому при наличии хотя бы одной НП, в которой участвуют оба автора. Так, N^{ca} можно рассматривать с точки зрения распределения степеней узлов сети (см. [4]), ранжирования авторов (см. [5]), выявления сообществ (см. [6]), динамики развития научной деятельности (см. [7, 8]). Результаты анализа N^{ca} позволяют оперативно получить достаточно полное и обоснованное представление о совместной научной деятельности в анализируемой области.

Начиная с 2000 гг. традиционное представление N^{ca} на основе бинарных отношений сменилось на использование структур, отражающих множественные отношения между авторами. Так, в работе [9] утверждается, что соавторы НП имеют m -арные отношения друг с другом, в том смысле, что они прикреплены к публикации, где m — число соавторов, образующих группу. В качестве варианта представления, учитывающего m -арные отношения между авторами, рассматривается двудольный граф, в котором одно дольное множество представляет авторов, другое — публикации, подготовленные этими авторами, что позволяет использовать аппарат теории графов. Однако при этом теряется «однородность» определения узлов, следовательно, усложняется изучение таких топологических свойств, как связность и кластеризация. Поэтому в работе [10] для представления КС предлагается использовать гиперграф — обобщение графа [11] и называть его *гиперсетью*. Ребра гиперграфа могут связывать группы узлов любой мощности. Работа [10] является одной из первых публикацией, в которых предлагается представлять N^{ca} с помощью гиперграфа, вершины которого соответствуют авторам, ребра — группам авторов, имеющим совместные НП. В работе [12] отмечено, что такой подход позволяет сфокусироваться исключительно на авторах. Если же принимать во внимание как ученых, так и их НП, то следует рассматривать одну научную статью как определяющую одно отношение между соавторами.

Примеры анализа структуры научного соавторства, основанные на моделях с множественными отношениями, можно найти в работе [9], в которой исследуются параметры N^{ca} , представленной случайным двудольным графом аффилиации; в работе [13], описывающей технологии визуализации гиперграфов, обеспечивающие минимальную потерю информации; а также в работах [10, 14], определяющих специальные меры, позволяющие выявить важность авторов в рамках представления сетей соавторства гиперграфами.

В последнее десятилетие изучение реальных КС привлекает значительное внимание со стороны исследователей, представляющих различные научные дисциплины. Одна из основных тем — изучение топологических особенностей КС значительного размера. В данной работе изложен метод построения сети научного соавторства, основанный на реальных библиографических данных, извлеченных из БД RePEc, содержащей сведения о приблизительно $3,8 \times 10^6$ НП, включая названия и краткие сведения об авторах. На основе этой информации построена гиперсеть научного соавторства H^{ca} и измерены ее параметры. Полученные результаты, сопоставленные с изложенными в работах [15, 16], расширяют представление о научном соавторстве в исследуемой информационной среде.

1. Гиперграф

1.1. Основные определения. *Гиперграф* $H = (V, E)$ [11] на конечном множестве $V = \{v_1, v_2, \dots, v_n\}$ определяется семейством $E = (E_1, E_2, \dots, E_m)$ подмножеств множества V , таких что:

- а) $E_i \neq \emptyset, i = 1, \dots, m$;
- б) $\cup_{i=1}^m E_i = V$;
- в) $(\forall v_j \in V \exists E_i) v_j \in E_i, j = 1, \dots, n$.

Элементы v_1, v_2, \dots, v_n называются *вершинами* H , а E_1, E_2, \dots, E_m — *ребрами*, или *гиперребрами*. Заметим, что в современной литературе (см., например, [17]) используется определение гиперграфа, в котором ограничения а) — в) отсутствуют. Число вершин H называется *порядком* гиперграфа и обозначается $n(H)$ либо n . Число ребер называется *размером* и обозначается $m(H)$ либо m . *Ранг* гиперграфа $r(H)$ определяется как $r(H) = \max_j |E_j|$. Гиперграф $H = (V, E)$ называется *простым*, если выполняется условие $E_i \subseteq E_j \Rightarrow i = j$. *Мультигиперграфом* называют гиперграф с повторяющимися ребрами, т. е. существуют хотя бы два ребра $E_i, E_j, i \neq j$, такие что $E_i = E_j$.

Если $v_i \in E_j$, то ребро E_j называют *инцидентным* вершине v_i . Вершины $v_i, v_k \in V$ называются *смежными* (обозначаются $v_i \sim v_k$), если существует ребро $E_j \in E$, такое что $v_i \in E_j$ и $v_k \in E_j$ [10]. Таким образом, все ребра, содержащие вершину, инцидентны данной вершине, а все вершины, входящие в по крайней мере одно общее ребро, являются смежными. *Степенью вершины* $\deg(v_i)$ называют число гиперребер, содержащих вершину v_i [11]. *Степенью ребра* $\deg(E_i)$ — его мощность (т. е. число вершин в ребре).

Двойственным гиперграфу $H = (V, E)$ на множестве V называется гиперграф $H^* = (V^*, E^*)$, такой что его вершины соответствуют ребрам в H , а гиперребра H^* соответствуют вершинам H , с отношением инцидентности, которое связывает каждую вершину с гиперребрами H , в которые вершина входит. Формально:

$$\begin{cases} V^* = E; \\ E_i^* = \{E_j | v_i \in E_j \text{ в } H\}. \end{cases}$$

1.2. Матричное представление. Гиперграф $H = (V, E)$ может быть представлен *матрицей инцидентности* $C(H) = (c_{ij})$, столбцы которой соответствуют ребрам, строки — вершинам:

$$c_{ij} = \begin{cases} 1, & \text{если } v_i \in E_j; \\ 0, & \text{если } v_i \notin E_j. \end{cases} \quad (1)$$

В терминах $C(H)$ степенью вершины является сумма элементов соответствующей строки матрицы инцидентности:

$$\deg(v_i) = \sum_j c_{ij}. \quad (2)$$

Матрица инцидентности гиперграфа H^* может быть вычислена как $C^* = C^T$, таким образом, $(H^*)^* = H$. Все понятия, относящиеся к гиперграфу, верны для двойственного гиперграфа. Заметим, что гиперграф, двойственный гиперграфу без повторяющихся ребер, может иметь повторяющиеся ребра, а также может оказаться обыкновенным графом (примеры см. в [17, 18]).

Матрица смежности гиперграфа $A(H) = (a_{ij})$ — это квадратная матрица, элемент a_{ij} — это число ребер, содержащих одновременно вершины v_i и v_j , диагональные элементы матрицы A равны нулю:

$$\begin{aligned} (\forall v_i \in V) a_{ii} &= 0; \\ (\forall v_i, v_j \in V, i \neq j) a_{ij} &= |\{E_k \in E : \{v_i, v_j\} \subseteq E_k\}|. \end{aligned} \quad (3)$$

Матрица $A(H)$ может быть получена из матрицы $C(H)$ следующим образом:

$$A(H) = C \cdot C^T - D_v,$$

где C^T — транспонированная матрица инцидентности, D_v — матрица, диагональные элементы которой равны степеням соответствующих вершин гиперграфа: $d_{ii} = \deg(v_i)$.

На основе гиперграфа можно построить графы, отражающие парные отношения между вершинами. Так, матрица $A(H)$ может рассматриваться как матрица смежности мультиграфа (взвешенного графа) $G(H)$, называемого *ассоциированным* графом гиперграфа $H = (V, E)$ [19].

Построим граф $[H]_2 = (V, E_2)$, называемый *2-секцией* гиперграфа $H = (V, E)$, такой что множество его вершин совпадает с V и две его вершины являются смежными тогда и только тогда, когда они обе принадлежат одному ребру в H , т. е.:

$$[H]_2 = (V, E_2), \text{ где } \{v_i, v_j\} \in E_2 \Leftrightarrow E(v_i) \cap E(v_j) \neq \emptyset,$$

где $E(v)$ — множество всех ребер, содержащих вершину v .

В монографии [11] понятие *2-секции* определено для простого гиперграфа, обобщение *2-секции* приведено в монографии [18]. Результатом обобщения является мультиграф, соответствующий графу $G(H)$.

Построим квадратную матрицу A' путем замены в матрице смежности A всех ненулевых элементов на единицу:

$$\begin{aligned} (\forall v_i \in V) a'_{ii} &= 0; \\ (\forall v_i, v_j \in V, i \neq j) a'_{ij} &= \begin{cases} 1, & \text{если } a_{ij} \neq 0; \\ 0, & \text{если } a_{ij} = 0; \end{cases} \end{aligned} \quad (4)$$

Матрицу A' можно рассматривать как матрицу смежности графа $[H]_2$, являющегося *2-секцией* гиперграфа в определении, приведенном в работе [20].

Кликой в неориентированном графе называется подмножество вершин, каждые две из которых соединены ребром графа. Гиперграф H , в котором гиперребра являются максимальными (по включению) кликами $[H]_2$, называется *конформным*. Если гиперграф не конформный, то при переходе от гиперграфа к графу парных отношений не все структуры графа возникают на основании структуры гиперграфа. Необходимые и достаточные условия конформности приведены в монографиях [11, 17].

Утверждение 1 (теорема Гилмора). Гиперграф конформный тогда и только тогда, когда для любых попарно пересекающихся трех гиперребер E_1, E_2, E_3 существует ребро E_4 , такое что выполняется условие $(E_1 \cap E_2) \cup (E_1 \cap E_3) \cup (E_2 \cap E_3) \subseteq E_4$ [17].

Гиперграф $H_w = (V, E, w)$ называется *взвешенным*, если определена функция $w: E \rightarrow \mathbb{R}$, сопоставляющая с каждым ребром $E_i \in E$ его вес $w(E_i)$ [21]. Если C — матрица инцидентности H_w , то *взвешенная степень* гиперграфа определяется формулой

$$d^w(v_i) = \sum_j c_{ij} \cdot w_j, \quad (5)$$

где $w_j = w(E_j)$.

Пусть $W = \text{diag}(w_j)$ — диагональная $m \times m$ матрица весов ребер, $w_{jj} = w(E_j)$, а $D_w = \text{diag}(d_w(v_i))$ — диагональная $n \times n$ матрица взвешенных степеней вершин. Матрица смежности взвешенного гиперграфа H_w , согласно [21], определяется формулой

$$A_w = C \cdot W \cdot C^T - D_w. \quad (6)$$

1.3. *Маршруты и пути.* Маршрут длиной l в гиперграфе $H = (V, E)$ определяется как последовательность вершин $(v_1, v_2, \dots, v_{l+1})$ (не обязательно различных), таких что для любого $i = 1, 2, \dots, l$ существует ребро, содержащее v_i и v_{i+1} ; маршрут *замкнутый*, если $v_1 = v_{l+1}$. *Путь* — это маршрут, у которого все ребра и вершины различны [10]. *Дистанция* между вершинами — длина кратчайшего пути между ними.

Согласно [20], гиперграф *связный*, если существует путь между любыми двумя вершинами, а *компонента связности* гиперграфа $H = (V, E)$ определяется как максимальное подмножество вершин $V' \subseteq V$, для которых существует путь между любыми двумя вершинами этого подмножества. В работе [22] компонента связности определена как множество вершин V' вместе с ребрами, инцидентными этим вершинам.

1.4. *Двудольные графы.* Пусть $r \geq 2$ натуральное число. Граф $G = (V, E)$ называется r -дольным, если V допускает такое разбиение на r подмножеств, при котором вершины каждого ребра принадлежат разным подмножествам [23]. При $r = 2$ такие графы называются *двудольными*.

Каждый конечный гиперграф $H = (V, E)$ может быть представлен двудольным графом $K(H) = (V \cup V', E_K)$, называемым *представлением Кеннига* [24], или *графом инцидентности*. Двудольный граф может быть получен путем сопоставления с каждым гиперребром $E_j \in E$ дополнительной вершины v_{e_j} и определения дольного множества $V' = \{v_{e_j} : E_j \in E\}$, при этом ребро между $v \in V$ и $v_{e_j} \in V'$ существует тогда и только тогда, когда $v \in E_j$. И наоборот, если имеется двудольный граф $G = (U, V, E)$ без изолированных вершин, то соответствующий ему гиперграф $H = (U, V)$ имеет множество вершин U и множество ребер V , такие что ребро $v = \{u \in U | (u, v) \in E\}$. Тем не менее, это разные структурные образования: гиперграф является расширением определения графа, а двудольный граф — частный случай графа (подробнее см. [20]).

Утверждение 2. Пусть гиперграф $H = (V, E)$ не имеет пустых ребер. Гиперграф H связный тогда и только тогда, когда связным является его граф инцидентности $K(H)$ [22].

Утверждение 3. Существует взаимно однозначное соответствие между компонентами гиперграфа и компонентами его графа инцидентности [22].

Представление гиперграфов двудольными графами позволяет применять аналитические методы исследования, используемые для анализа сетей, построенных на основе бинарных отношений [25].

2. Моделирование множественных отношений между авторами. Гиперграфы являются естественным средством моделирования связей между группами соавторов. Обозначим $P = \{p_1, p_2, \dots, p_m\}$ множество НП, $S = \{s_1, s_2, \dots, s_n\}$ — множество их авторов. Предполагаем, что P содержит только НП, подготовленные двумя и более авторами, т. е. в конструируемом гиперграфе не будет ребер, состоящих из одной вершины. Представим два метода построения гиперграфа соавторства.

2.1. Определим гиперграф $H_1 = (V, E_1)$, такой что множество S отображается на множество вершин V , а множество P — на множество ребер E_1 , причем если НП p_i подготовлена именно авторами v_1, v_2, \dots, v_k , то $E_i = \{v_1, v_2, \dots, v_k\}$ является ребром, $E_i \in E_1$. Число

ребер $m_1 = |E_1|$ — это число публикаций $|P|$. При таком методе гиперграф становится мультигиперграфом, если группа авторов s_1, s_2, \dots, s_k подготовила совместно несколько НП. Подобный метод рассматривается в работе [10]. В работе [9] определен двудольный граф сотрудничества, который может рассматриваться как граф инцидентности $K(H_1)$ гиперграфа H_1 . В этом случае принимаются во внимание как группы авторов, так и публикации. Полученный мультигиперграф по аналогии с понятием мультиграфа является мультигиперграфом с собственной идентификацией каждого ребра, т. е. мультиребра, состоящие из одного и того же множества вершин, независимы. Так, в мультиграфе с независимыми ребрами каждое ребро может иметь свой вес.

2.2. Если изучать исключительно группы авторов, публикующихся совместно, то определяем взвешенный гиперграф $H_2 = (V, E_2, w)$, множество S отображается на V , $E_i = \{v_1, v_2, \dots, v_k\}$ является ребром, $E_i \in E_2$, если существует хотя бы одна НП, совместно опубликованная авторами v_1, v_2, \dots, v_k . Вес ребра — число НП, опубликованных совместно данными k авторами [6]. Число ребер $m_2 = |E_2|$ — число групп авторов. В работе [14] веса присваиваются также вершинам, вес вершины — это общее число НП, подготовленных автором.

Анализ соавторства по методу 2.1 предполагает, что матрица смежности $A(H_1)$ определяется согласно (3), тогда как при анализе по методу 2.2 матрица $A(H_2)$ определяется согласно (6). Важно отметить, что $A(H_1) = A(H_2)$, следовательно, совпадают ассоциированные графы гиперграфов: $G(H_1) = G(H_2)$. Однако двудольные графы инцидентности $K(H_1)$ и $K(H_2)$ различны.

3. Измерение параметров. При анализе гиперграфов зачастую используются расширения инструментария теории графов. Однако в ряде случаев возникают серьезные теоретические препятствия. В первую очередь это относится к спектральной теории гиперграфов [26, 27], поскольку матрица смежности недостаточна для кодирования отношений смежности в гиперграфе, в котором ребра задаются множествами вершин. Также исследование гиперграфов требует специального рассмотрения таких понятий как модульность и выявление сообществ [28, 29].

Одним из способов исследования является преобразование гиперграфа в граф, например в *2-секцию* или граф инцидентности. Так, гиперграфы изоморфны своим графам инцидентности, поэтому ряд аналитических методов исследования двудольных графов естественным образом распространяется на гиперграфы. Однако не все понятия, применимые к гиперграфу, свойственны двудольному графу. Например, понятие коэффициента кластеризации для гиперграфа неприменимо к двудольным графам ввиду отсутствия треугольных структур (см. [30]). В настоящей работе используется представление гиперграфа в виде двудольного графа инцидентности для вычисления ряда параметров.

3.1. *Параметры и свойства.* Используя переход к двудольному графу инцидентности $K(H) = (V, V', E_K)$, можно определить связность H (число, состав и размер компонент, см. утверждения 2, 3) и вычислить следующие параметры (см. 1.4):

- а) размер и порядок гиперграфа H : $n(H) = |V|, m(H) = |V'|$;
- б) среднее расстояние в H (среднее расстояние в $K(H)$ между вершинами дольного множества V , деленное на два);
- в) распределение степеней вершин и ребер H , максимальную и среднюю степень на основе вершин каждого из дольных множеств $K(H)$.

На основе $K(H)$ невозможно выявить свойства простоты и конформности H . Непростой гиперграф, включающий вложенные ребра, отображается в простой граф $K(H)$.

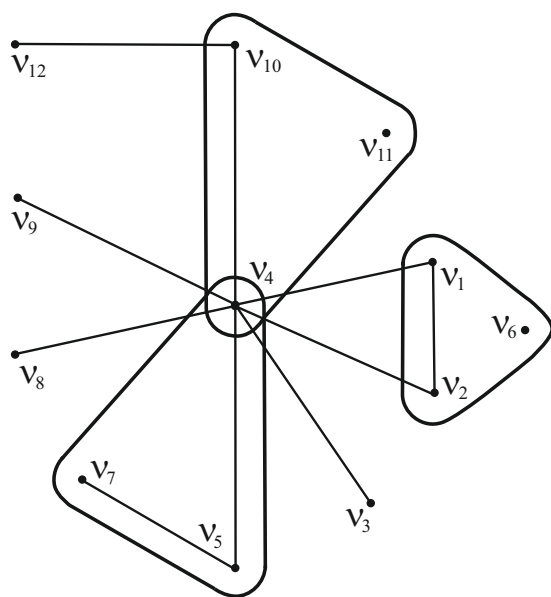
Рис. Компонента H_{12}

Таблица 1.

Вершины и ребра H_{12}

| Гиперребро | Вершины гиперребра | Примечания | |
|-------------------|-----------------------|------------|------|
| e_1 | v_1, v_2, v_6 | 13 НП | |
| $e_2 - e_{14}$ | v_1, v_2 | | |
| e_{15} | v_4, v_1 | | |
| e_{16} | v_4, v_2 | | |
| e_{17} | v_4, v_3 | | |
| e_{18} | v_4, v_5 | | |
| e_{19} | v_4, v_5, v_7 | | |
| e_{20} | v_4, v_8 | | |
| e_{21} | v_4, v_{10} | | |
| e_{22} | v_4, v_{10}, v_{11} | | |
| $e_{23} - e_{25}$ | v_5, v_7 | | 3 НП |
| e_{26} | v_4, v_9 | | |
| e_{27} | v_{10}, v_{12} | | |

Таблица 2.

Параметры H_{12}

| Параметр | Значение |
|----------------------------------------------|----------|
| Размер $m(H_{12})$ | 27 |
| Порядок $n(H_{12})$ | 12 |
| Ранг $r(H_{12})$ | 3 |
| Максимальная степень вершин $\max deg(v_i)$ | 15 |
| Средняя степень вершин $\text{avr} deg(v_i)$ | 4,75 |
| Среднее расстояние $L(H_{12})$ | 2 |

Кроме того, граф инцидентности непростого гиперграфа с помеченными мультиребрами также является простым. Понятие конформности не определено для двудольного графа.

3.2. *Исходные данные.* В качестве исходных данных рассматривается множество НП, проиндексированных в БД RePEc на момент извлечения (2020.01.31). Проблема распознавания индивидуальных авторов на основании текста НП во многом зависит от того, насколько хорошо документирована база данных. В нашем случае при отборе авторов и НП была использована информация, содержащаяся в профилях авторов. Процедура фильтрации «сырых» данных представлена в работе [15]. В результате было выделено 91 113 НП, подготовленных в соавторстве, и 32 434 автора этих НП. Для построения гиперграфа соавторства $H^{ca} = (V, E)$ используется метод, указанный в п. 2.1.

В результате исследования $K(H^{ca})$ выявлено, что H^{ca} не является связным. Для компоненты H_{12} , состоящей из 12 вершин (авторов) и 27 гиперребер (НП), вычислим значения параметров и определим свойства H_{12} (рисунок, табл. 1).

Параметры H_{12} приведены в табл. 2.

Таблица 3.

Основные параметры H^{ca}

| Параметр | Значение |
|---------------------------------------------------------------------|-------------------------------------------------------------|
| Размер $m(H^{ca})$ | 91 113 |
| Порядок $n(H^{ca})$ | 32 434 |
| Ранг $r(H^{ca})$ (см. Замечание) | 17 |
| Число вершин максимальной компоненты | 29 270 |
| Средняя степень $avrdeg(E_i)$ | 2,3 |
| Средняя степень $avrdeg(v_i)$ | 2,8 |
| Распределение степеней вершин $\deg(v_i)$ следует степенному закону | $Px \approx x^{-\gamma}$, $x_{\min} = 6, \gamma = 1,64$ |

Свойства H_{12} :

а) H_{12} не является простым гиперграфом, поскольку имеет повторяющиеся и вложенные гиперребра, например $e_{18} \subset e_{19}$ (имеется НП, подготовленная авторами v_4, v_5 , и другая НП, подготовленная авторами v_4, v_5, v_7);

б) H_{12} не является конформным, так как попарно пересекающиеся ребра e_2, e_{15}, e_{16} не удовлетворяют теореме Гилмора (см. утверждение 1).

Основные параметры H^{ca} приведены в табл. 3.

Свойства «не простоты» и «не конформности» H^{ca} вытекают из свойств компоненты H_{12} .

Замечание. В нашем случае основным содержанием БД RePEc являются НП в области экономики, где публикации с 2–3 соавторами являются нормой, максимальное число соавторов равно 17. В то же время в области математики большинство НП выполнены одним автором, а НП с тремя и более соавторами являются редкостью. Крайний случай демонстрируют работы в области экспериментальной физики. Для примера приведем НП “The ATLAS Experiment at the CERN Large Hadron Collider”, подготовленную 2927 соавторами, аффилированными в 277 организациях (Journal of Instrumentation. 2008. Vol. 3, iss. 8. DOI:10.1088/1748-0221/3/08/S08003).

Заключение. Перечислим свойства гиперграфа H^{ca} :

а) не является связным, связность и число компонент определены путем анализа графа инцидентности $K(H)$ (см. утверждения 2, 3);

б) не является простым, имеет повторяющиеся и вложенные ребра, например $e_{18} \subset e_{19}$ (см. рис. 1);

в) не является конформным: он содержит по крайней мере три ребра таких, что не существует гиперребра, включающего объединение их попарных пересечений, например это ребра e_2, e_{15}, e_{16} (см. Утверждение 1).

Следует отметить, что информация о соавторстве, полученная при исследовании сетей N^{ca} и N_T^{ca} [15, 16], представленных графами, может быть получена из гиперграфа H^{ca} . Поясним это утверждение.

Во-первых, информация о соавторстве, приведенная в табл. 1 в работе [15], может быть извлечена из параметров гиперграфа. Заметим, что гиперграф строился без учета индивидуальных работ авторов, поэтому $P = P_+$. Если снять ограничение, то появляются гиперребра, имеющие в составе одну вершину, число таких ребер равно P_1 . Максимальное число соавторов, среднее число соавторов НП и среднее число НП автора — это ранг и средние степени вершин и ребер гиперграфа H^{ca} .

Во-вторых, параметры невзвешенной сети N^{ca} можно исследовать, если перейти к представлению гиперребер H^{ca} в виде клик, т. е. рассматривать параметры 2-секции H^{ca} — графа $[H]_2$. Заметим, что если определить степень вершины в гиперграфе как число смежных вершин, то получим степень вершины в сети N^{ca} .

В-третьих, параметры взвешенной сети $N_T^{ca} = (V^{ca}, E^{ca}, W_T)$, рассмотренной в работе [16], могут быть проанализированы с использованием ассоциированного графа $G(H^{ca})$, соответствующего N_T^{ca} , поскольку в матрице весов $W_T = (w_{ij})$ элемент w_{ij} равен числу соавторских связей между учеными v_i и v_j , т. е. числу НП, авторами которых являются и v_i , и v_j .

Список литературы

1. BOCCALETTI S., LATORA V., MORENO Y., CHAVEZ M., HWANG D. U. Complex networks: Structure and dynamics // Phys. Rep. 2006. V. 424, iss. 4–5. P. 175–308. DOI: 10.1016/j.physrep.2005.10.009.
2. BATTISTON F., CENCETTI G., IACOPINI I., LATORA V., LUCAS M., PATANIA A. YOUNG J-G., PETRI G. Networks beyond pairwise interactions: Structure and dynamics // Phys. Rep. 2020. V. 874. P. 1–92. DOI: 10.1016/j.physrep.2020.05.004.
3. ЩЕРБАКОВА Н. Г. Моделирование групповых взаимодействий комплексных систем. Обзор // Пробл. информ. 2022. № 3. С. 24–45.
4. NEWMAN M. E. J. Scientific collaboration networks. I. Network construction and fundamental results // Phys Rev. E, 64(1), 016131. DOI: 10.1103/PhysRevE.64.016131.
5. NEWMAN M. E. J. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality // Phys. Rev. E, 64(1), 016132. DOI: 10.1103/PhysRevE.64.016132.
6. SAVIĆ M., IVANOVIĆ M., RADOVANOVIĆ M., OGNJANOVIĆ Z., PEJOVIĆ A. Exploratory analysis of communities in co-authorship networks: A case study // Intern. Conf. on ICT Innovations, Springer, 2019. P. 55–64.
7. BARABÁSI A. L., JEONG H., NEDA Z., RAVASZ E., SCHUBERT A., VICSEK T. Evolution of the social network of scientific collaborations // Physica A. 2002. V. 311. P. 590–614. DOI: 10.48550/arXiv.cond-mat/0104162.
8. UDDIN S., HOSSAIN L., ABBASI A., RASMUSSEN K. Trend and efficiency analysis of co-authorship network // Scientometrics. 2012. V. 90, No. 2. P. 687–699. DOI: 10.1007/s11192-011-0511-x.
9. NEWMAN M. E. J., STROGATZ S. H., WATTS D. J. Random graphs with arbitrary degree distributions and their applications // Phys. Rev. E 64, 026118. 2001. DOI: 10.1103/PhysRevE.64.026118.
10. ESTRADA E., RODRIGUEZ-VELAZQUEZ J. A. Complex networks as hypergraphs // Arxiv: physics/0505137, 2005. DOI: 10.1016/j.physa.2005.12.002.
11. BERGE C. Hypergraphs. Amsterdam; N. Y.; Oxford; Tokyo: North-Holland, 1989.
12. TORRES L., BLEVINS A. S., BASSET D., ELIASSI-RAD T. The why, how, and when of representations for complex systems // SIAM Rev. 2021. V. 63, No 3. P. 415–485. DOI: 10.1137/20M1355896.
13. OUVREARD X., LE GOFF X-M., MARCHAND-MAILLET S. Networks of collaborations: Hypergraph modeling and visualization // ArXiv: 1809.00164v1. DOI: 10.48550/arXiv.1809.00164.
14. HAN Y., ZHOU B., PEI J., JIA Y. Understanding importance of collaborations in coauthorship networks: A supportiveness analysis approach // Proc. 2009 SIAM Intern. Conf. on Data Mining. 2009. P. 1112–1123. DOI: 10.1137/1.9781611972795.95.

15. БРЕДИХИН С. В., ЛЯПУНОВ В. М., ЩЕРБАКОВА Н. Г. Структура и параметры невзвешенной сети соавторства на основе данных БД RePEc // Пробл. информ. 2021. № 3. С. 56–67. DOI: 10.24411/2073-0667-2021-3-56-57.
16. БРЕДИХИН С. В., ЛЯПУНОВ В. М., ЩЕРБАКОВА Н. Г. Ранжирование узлов взвешенной сети соавторства: анализ данных БД RePEc // Пробл. информ. 2021. № 4. С. 5–15. DOI: 10.24412/2073-0667-2021-4-67-83.
17. VOLOSHIN V. I. Introduction to graph and hypergraph theory. N. Y.: Nova Science Publishers, Inc., 2009.
18. BRETTO A. Hypergraph theory: An introduction. Heidelberg: Springer Intern. Publishing, 2013. DOI: 10.1007/978-3-319-00080-0.
19. MARTINEZ M. G., STARK H. M., TERRAS A. A. Some Ramanujan hypergraphs associated to $GL(n, \mathbb{F}_q)$ // Proc. Am. Math. Soc. 2001. V. 129, P. 1623–1629. S. 0002-9939(00)05965-7.
20. OUVARD X. Hypergraphs: An introduction and review // Arxiv: 2002.05014v2, 2020. DOI: 10.48550/arXiv.2002.05014.
21. ZHOU D., HUANG J., SCHÖKOPF B. Learning with hypergraphs: Clustering, classification, and embedding // Proc. 19th Internat. Conf. on Neural Inform. Proc. Syst. 2007. P. 1601–1608. DOI: 10.7551/mitpress/7503.003.0205.
22. ВАХМАНИАН М. А., САЈНА М. Connection and separation in hypergraphs // Theory and Appl. of Graphs. 2015. V. 2, iss. 2. DOI:10.20429/tag.2015.020205.
23. ДИСТЕЛЬ Р. Теория графов. Пер. с англ. Новосибирск: Изд-во Ин-та математики, 2002. 336 с.
24. ЗЫКОВ А. А. Гиперграфы // Успехи матем. наук. 1974. Т. 29, вып. 6. С. 89–156.
25. BORGATTI S. P., EVERETT M. G. Network analysis of 2-mode data // Social networks. 1997. V. 19. P. 243–269. DOI: 10.1016/S0378-8733(96)00301-2.
26. COOPER J., DUTLE A. Spectra of uniform hypergraphs // Lin. Algebra and Its Appl. 2012. V. 436. P. 3268–3292. DOI: 10.48550/arXiv.1106.4856.
27. BANERJEE A., CHAR A., MONDAL B. Spectra of general hypergraphs // Lin. Algebra and Its Appl. 2017. V. 518. P. 14–30. DOI: 10.1016/j.laa.2016.12.022.
28. KUMAR T., VAIDYANATHAN S., ANANTHAPADMANABHAN H. Hypergraph clustering: A modularity maximization approach // ArXiv: 1812.10869[cs.G]. DOI: 10.48550/arXiv.1812.10869.
29. KAMIŃSKI B., POULIN V., PRALAT P., SZUFEL P., THÈBERGE F. Clustering via hypergraph modularity // PLoS ONE. 2019. V. 14(11), e0224307. DOI: 10.1371/journal.pone.0224307.
30. ZHOU V., NAKHLEN L. Properties of metabolic graphs: biological organization or representation artifacts? // BMC Bioinformatics. 2011. V. 12, 132. DOI: 10.1186/1471-2105-12-132.



Бредихин Сергей Всеволодович — канд. техн. наук, зав. лабораторией Ин-та вычислительной математики и математической геофизики СО РАН; e-mail: bred@nsc.ru;

Сергей Бредихин окончил механико-математический факультет Новосибирского государственного университета в 1968 г. С 1968 г. — сотрудник Института автоматизации и электрометрии СО РАН. Кандидат технических наук с 1983 г. С 1988 г. — заведующий

Лабораторией прикладных систем Института вычислительной математики и математической геофизики СО РАН. Являлся техническим директором проекта «Сеть Интернет Новосибирского Научного Центра». Лауреат государственной премии по науке и технике 2012 г. В сфере его научных интересов — измерение и анализ сетей распределенных информационных структур. Автор и соавтор более 110 работ и двух монографий: «Методы библиометрии и рынок электронной научной периодики», «Анализ цитирования в библиометрии».

Sergey Bredikhin graduated from Novosibirsk State University in 1968 (faculty of Mechanics and Mathematics). In 1968 he became an employee of Institute of Automation and Electrometry SB RAS. In 1983 he received PhD degree in Engineering Science. Since 1988 he is the head of Applied Systems laboratory of Institute of Computational Mathematics and Mathematical Geophysics SB RAS. He was the technical manager of «Akademgorodok Internet Project». He is the state prize winner in science and engineering (2012). Sphere of his scientific interests — the measurement and analysis of networks of the distributed information structures. He is the author and co-author of more than 110 works and two monographs: «Metody bibliometrii i rynok elektronnoj nauchnoj periodiki», «Ansliz tsitirovaniya v bibliometrii».



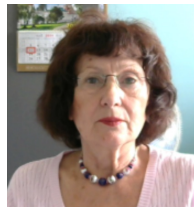
Ляпунов Виктор Михайлович — ведущий инженер Ин-та вычислительной математики и математической геофизики СО РАН; e-mail: vic@nsc.ru;

Виктор Ляпунов окончил механико-математический факультет Новосибирского государственного университета в

1978 году. В 1978 г. стал сотрудником Вычислительного Центра СО АН СССР, а с 1990 г. — сотрудником Института систем информатики СО АН СССР. С 2004 г. — ведущий инженер Института вычислительной математики и математической геофизики СО РАН. Занимается вопросами извлечения информации из баз данных и обработкой больших массивов данных. Соавтор более 10 работ в этой области.

Victor Lyapunov graduated from Novosibirsk State University in 1978 (faculty of Mechanics and Mathematics). In 1978, he became an employee of Computing Center of SB AS USSR, since 1990 — an employee of Institute of Informatics Systems SB RAS. Since 2004 he works as software engineer in Institute of Computational Mathematics and Mathematical Geophysics SB RAS. His current research interests include methods of information extracting from

databases and processing of large data sets. He is the co-author of more than 10 works in that area.



Щербакова Наталья Григорьевна — ст. науч. сотр. Ин-та вычислительной математики и математической геофизики СО РАН; e-mail: nata@nsc.ru.

Наталья Щербакова окончила Новосибирский государственный университет по специальности «Математическая лингвистика» в 1967 г. С 1967 г. работала в Институте математики СО РАН, затем в Институте автоматизации и электрометрии СО РАН в области создания программного обеспечения систем передачи данных. С 2000 г. — сотрудник Института вычислительной математики и математической геофизики СО РАН, где с 2002 г. занимает должность старшего научного сотрудника. Являлась участником проекта «Сеть Интернет Новосибирского Научного Центра», занималась вопросами мониторинга и анализа IP-сетей. Автор и соавтор более 40 работ, соавтор монографии «Анализ цитирования в библиометрии». Текущие интересы лежат в области исследования методов оценки научной деятельности на основе анализа цитирования научной литературы.

Natalia Shcherbakova graduated from Novosibirsk State University in 1967 (mathematical linguistics). Since 1967 she worked at Institute of Mathematics SB RAS, then at Institute of Automation and Electrometry SB RAS in the field of software design for data transmission systems. In 2000 — the employee of Institute of Computational Mathematics and Mathematical Geophysics SB RAS, since 2002 works as senior researcher. She is a member of «Akademgorodok Internet Project», dealt with software of monitoring and the analysis of IP networks. She is the author and co-author of more than 40 works, the co-author of the monograph «Ansliz tsitirovaniya v bibliometrii». The current research interests lie in the field of bibliometrics: methods of measuring of scientific activity on the base of citations.