

## CLUSTER ANALYSIS OF THE CITATION NETWORK OF SCIENTIFIC JOURNALS

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

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

---

In this work we analyze the structure of the journal citation network built on the basis of the bibliographic information extracted from the database *RePEc*. The network is represented as the weighted directed graph (digraph), the main component  $G$  has 1729 vertices (journals) and 135702 arcs (citations). In accordance with *M. Kessler* (1963) the network of bibliographic coupling was constructed that is represented as weighted undirected graph  $G^{bib}$  and in accordance with *G. Small* (1973), *I. V. Marshakova* (1973) - co-citation network was constructed that is represented as weighted undirected graph  $G^{coc}$ . The weights of edges are assigned accordingly vector space model *G. Salton*, *M. MacGill* (1983). The graphs  $G$ ,  $G^{bib}$  и  $G^{coc}$  are the objects of studying.

In the first part of the work we examine the problem of network connectivity via the adjacency relations between neighbors. The answer comes in two main flavors. One approach assesses the overall level of clustering in a network, and is called transitivity, see *S. Wasserman*, *K. Faust*, (1994). The global clustering coefficient is the fraction of closed triplets (subgraphs with three nodes and three edges) to all triplets (subgraphs with three nodes and two edges), see *M. Newman* (2002). A generalization to weighted networks was proposed by *T. Opsahl*, *P. Panzarasa* (2009). An alternative approach to connectivity was introduced in the work *D. Watts*, *S. Strogatz* (1998). A node clustering coefficient is defined as the fraction of number of actual ties among the neighbor nodes over possible ties between them. The network local clustering coefficient is defined as an average of clustering coefficients of nodes and is considered as one of the small-world parameters. For the weighted networks we use the generalization proposed in the work *A. Barrat* (2004). The results of local clustering measurements for  $G$ ,  $G^{bib}$  и  $G^{coc}$  are presented.

In the second part of the work we analyze the community structure of the weighted digraph  $G$  and weighted undirected graphs  $G^{bib}$  и  $G^{coc}$ . It will be remarked that the majority of clustering algorithms are designed for weighted undirected graphs. We examine applicability of algorithms *BTW* *M. Girvan*, *M. Newman* (2002), *WTR* *P. Pons*, *M. Latapy* (2005), *IMP* *M. Rosvall*, *C. Bergstrom* (2008), *MLO* *V. Blondel* (2008). For digraph  $G$  clustering we used algorithms *BTW* и *IMP*, and for undirected version of  $G$  we used the whole kit. In both cases *IMP* obtained the community structure that largely corresponds to the economic disciplines. For clustering  $G^{bib}$  and  $G^{coc}$  the algorithms *IMP*, *WTR* и *MLO* were applied. Measures for the similarity of partitions delivered by the algorithms were analyzed (*NMI*, *RAND*). The results of applying community detection algorithms to graphs  $G$ ,  $G^{bib}$  and  $G^{coc}$  are presented in the tables (3.1-3.4).

The conclusion contains comments to the results of the research. The approved tools give the basic insight about the structure of the bibliometric networks on study.

**Key words:** journal citation network, co-citation network, bibliographic coupling network, weighted directed graph, transitivity, weighted local clustering coefficient, community finding.

## References

1. RePEc. General principles. [Electron. Resource]. <http://repec.org/>.
2. BREDIKHIN S. V., LYAPUNOV V. M., SHCHERBAKOVA N. G. The structure of the citation network of scientific journals // *Problemy informatiki*. 2017. № 2. P. 38–52.
3. HARARY F. *Graph Theory*. Addison-Wesley, 1969.
4. KESSLER M. M. Bibliographic coupling between scientific papers // *American Documentation*. 1963. V. 14. P. 10–25.
5. SALTON G., MACGILL M. J. *Introduction to modern information retrieval*. N. Y.: McGraw-Hill, 1983.
6. SMALL H. Co-citation in the scientific literature: A new measure of the relationship between two documents // *J. of the American Society for Information Science*. 1973. V. 24. P. 265–269.
7. MARSHAKOVA I. system of document connections based on references // *Scientific and Technical Information Serial of VINITI*. 1973. V. 6, N 2. P. 3–8.
8. WATTS D. J., STROGATZ S. H. Collective dynamics of „small-world“ networks // *Nature*. 1998. V. 393. P. 440–442.
9. WASSERMAN S., FAUST K. *Social network analysis: Methods and applications*. Cambridge (ENG), New York: Cambridge University Press, 1994.
10. BRANDES U. *Network analysis*. Berlin, Heidelberg, New York: Springer, 2005.
11. NEWMAN M. E. J., STROGATZ S. H., WATTS D. J. Random graph models of social networks // *Proc. of the National Academy of Science of the USA*. 2002. V. 99. P. 2566–2572.
12. BOLLOBAS B., RIORDAN O. M. Mathematical results on scale-free random graphs. *Handbook of graphs and networks: From genome to Internet*. Weinheim, FRG: Wiley-VCH Verlag GmbH & Co. KGaA, 2002. P. 1–34.
13. BARRAT A., BARTHELEMY M., PASTOR-SATORRAS R., VESPIGNANI A. The architecture of complex weighted networks // *Proc. of the National Academy of Sciences*. 2004. V. 101, iss. 11. P. 3747–3752.
14. LOPEZ-FERNANDEZ L., ROBLES G., GONZALEZBARAHONA J. Applying social network analysis to the information in cvs repositories // *Proc. of the 1st Intl. workshop on mining software repositories (MSR2004)*. USA: Springer, 2004. P. 101–105.
15. ONNELA J.-P., SARAMI J., KERTZ J., KASKI K. Intensity and coherence of motifs in weighted complex networks // *Phys. Rev.* 2005. E 71 065103.
16. ZHANG B., HORVATH S. A general framework for weighted gene co-expression network analysis // *Statistical Applications in Genetics and Molecular Biology*. 2005. V. 4., N 17.
17. OPSAL T., PANZARASA P. Clustering in weighted networks // *Social networks*. 2009. V. 31. P. 155–163.
18. FORTUNATO S. Community detection in graphs // *Physics Reports*. 2010. V. 486. P. 75–174.
19. MALLIAROS F. D., VAZIRGIANNIS M. Clustering and community detection in directed networks: A survey // *Physics Reports*. 2013. V. 533, iss. 4. P. 95–142.
20. KANNAN R., VAMPALA S., VETTA A. On clustering — good, bad and spectral // *Foundations of Computer Science*. 2000. P. 367–378.
21. VAN DONGEN S. M. Graph clustering by flow simulation // *PhD thesis*. University of Utrecht. 2000.
22. NEWMAN M. E. J., GIRVAN M. Finding and evaluating community structure in networks // *Phys. Rev.* 2004. E 69 (2) 026113.
23. ARENAS A., DUCH J., FERNANDEZ A., GÓMEZ S. Size reduction of complex networks preserving modularity // *New J. Phys.* 2007. V. 9, N. 6. P. 176–190.
24. NEWMAN M. E. J. Analysis of weighted networks // *Phys. Rev. E* 70, 056131. 2004.

25. GÓMEZ S., JENSEN P., ARENAS A. Analysis of community structure in networks of correlated data // *Phys. Rev. E* 80, 016114.
26. FRED A. L. N., JAIN A. K. Robust data clustering // *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, Minneapolis, USA, June 16–22, 2003*. P. 128–136.
27. RAND W. M. Objective criteria for the evaluation of clustering methods // *J. Amer. Statistical Association*. 1971. V. 66, N. 336. P. 846–850.
28. GIRVAN M., NEWMAN M. E. J. Community structure in social and biological networks // *Proc. Nat. Acad. Sci. USA*. 2002. V. 99. P. 7821–7826.
29. FREEMAN L. C. A set of measures of centrality based upon betweenness // *Sociometry*. 1977. V. 40. P. 35–41.
30. ROSVALL M., BERGSTROM C. T. Maps of random walks on complex networks reveal community structure // *Proc. Natl. Acad. Sci. USA*. 2008. V. 105, N 4. P. 1118–1123.
31. RISSANEN J. Modeling by short data description // *Automatica*. 1978. V. 14. P. 465–471.
32. SHANNON C. E. A mathematical theory of communications // *Bell System Tech. J.* 1948. V. 27. P. 379–423.
33. PONS P., LATAPY M. Computing communities in large networks using random walks // *J. of Graph Algorithms and Applications*. 2006. V. 10, N 2. P. 191–218.
34. BLONDEL V., GUILLAUME J., LAMBIOTTE J., LEFEBVRE E. Fast unfolding of communities in large networks // *J. Stat. Mech.* 2008, P10008.
35. NEWMAN M. E. J. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality // *Phys. Rev. E* 64, 016132. 2001.
36. BRANDES U. On variants of shortest-path betweenness centrality and their generic computation // *Soc. Networks*. 2008. V. 30. P. 136–145.
37. Jel Classification System / EconLit subject descriptors. 2016. [Electron. resource]. <https://www.aeaweb.org/econlit/jelCodes.php?view=jel>.
38. BREDIKHIN S. V., LYAPUNOV V. M., SHCHERBAKOVA N. G. The Structure of the Citation Network of Scientific Journals // *Problemy informatiki*. 2016. N 3. P. 28–43.

# КЛАСТЕРНЫЙ АНАЛИЗ СЕТИ ЦИТИРОВАНИЯ НАУЧНЫХ ЖУРНАЛОВ

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

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

УДК 001.12+303.2

Изучается сеть цитирования научных журналов, представленная взвешенным ориентированным графом. Основное внимание сфокусировано на проблемах связности и выявлении модульной структуры сети. Рассмотрены методы анализа объектов сетевой структуры. На основе реальных библиографических данных, извлеченных из БД *RePEc* [1], построены главная связная сетевая компонента  $G$  и производные сети: коцитирования —  $G^{coc}$  и библиографического сочетания —  $G^{bib}$ . Для этих сетей измерены локальный и взвешенный коэффициенты кластеризации. Выявление модульности рассматривается как задача идентификации структурно эквивалентных вершин соответствующих графов. С применением алгоритмов *BTW*, *IMP*, *WTR* и *MLO* выполнен кластерный анализ компоненты  $G$  и производных сетей. Результаты представлены в виде рисунка и таблиц. Сравнение результатов осуществлено с помощью индексов согласованности *NMI* и *RAND*.

**Ключевые слова:** сеть цитирования научных журналов, сети коцитирования и библиографического сочетания, взвешенный ориентированный граф, локальный взвешенный коэффициент кластеризации, выявление сообществ.

**1. Обозначения и определения.** Эта работа является продолжением [2]. Напомним основные сведения, которые потребуются в дальнейшем.

1.1. *Сеть цитирования журналов (СЦЖ).* На множестве журналов  $J = \{J_1, J_2, \dots, J_n\}$  задано отношение цитирования  $R$ :

$$J_k R J_l \equiv J_k \text{ цитирует } J_l. \quad (1.1)$$

СЦЖ представляет оргграф  $G = (V, E)$ , в котором журналы соответствуют вершинам  $V = \{v_1, v_2, \dots, v_n\}$ , а цитирования — дугам  $E \subseteq V \times V$ ,  $e = (v_i, v_j) \in E$ , если выполняется (1.1). Если  $i = j$ , то отношение (1.1) называется самоцитированием. На множестве дуг определена функция  $w : E \rightarrow \mathbb{N}^+$ , такая что  $w(e) = w_{ij}$  равна числу цитирований, полученных журналом  $j$  от журнала  $i$ . Матрица цитирований  $A$  графа  $G$  имеет размер  $|V|$  и содержит элементы:

$$A(i, j) = \begin{cases} 1, & \text{если } i R j, \\ 0, & \text{в противном случае.} \end{cases}$$

Матрица весов  $W$  дуг графа  $G$  имеет размер  $|V|$  и содержит элементы:

$$W(i,j) = \begin{cases} w_{i,j}, & \text{если } iRj, \\ 0, & \text{в противном случае.} \end{cases}$$

СЦЖ соответствует слабо связный [3] взвешенный оргграф  $G = (V,E), |V| = 1729, |E| = 135702$  (без учета самоцитирований и изолированных вершин). Максимальная сильно связная компонента  $G$  имеет 1278 вершин.

1.2. *Сеть библиографического сочетания журналов* (СБСЖ) построена на основе СЦЖ. Метод „библиографического сочетания“ [4] был распространен на множество журналов. Говорят, что журналы  $i$  и  $j$  находятся в состоянии библиографического сочетания, если существует журнал  $k$ , на который ссылаются  $i$  и  $j$ :

$$iR^{bib}j \equiv (\exists k) iRk \& jRk. \quad (1.2)$$

СБСЖ соответствует взвешенный граф  $G^{bib}$ , множество вершин которого совпадает с  $V$ . Отношения  $R^{bib}$  между вершинами выступают в роли неориентированных связей (ребер)  $E \subseteq V \times V, e = (i,j) \in E$ , если имеет место  $iR^{bib}j$ . Журнал  $i$  представляется вектором (строкой) матрицы  $W$ . Для определения веса ребер графа  $G^{bib}$  используется векторная модель [5], а степень подобия  $bib(i,j)$  вычисляется следующим образом:

$$bib(i,j) = \frac{\sum_k w_{ik}w_{jk}}{\sqrt{\sum_k w_{ik}^2} \sqrt{\sum_k w_{jk}^2}} \quad (1.3)$$

Матрица смежности  $G^{bib}$ , соответствующего СБСЖ, обозначается  $W^{bib}$ . В качестве графа  $G^{bib} = (V^{bib}, E^{bib})$  рассматривается его максимальная связная компонента,  $|V^{bib}| = 1432$  (остальные 297 вершин являются одиночными),  $|E^{bib}| = 844476$  ребер.

1.3. *Сеть коцитирования журналов* (СКЦЖ) также построена на основе СЦЖ. Для этого на множество журналов был распространен метод „коцитирования“ [6, 7]. Журналы  $i$  и  $j$  находятся в отношении коцитирования  $R^{coc}$ , если существует журнал  $k$ , содержащий ссылки на журналы  $i$  и  $j$ , т. е.

$$iR^{coc}j \equiv (\exists k) kRi \& kRj. \quad (1.4)$$

СКЦЖ соответствует взвешенный граф  $G^{coc}$ , множество вершин которого совпадает с  $V$ . Отношения  $R^{coc}$  между вершинами выступают в роли неориентированных связей (ребер)  $E \subseteq V \times V, e = (i,j) \in E$ , если имеет место  $iR^{coc}j$ . Журнал  $i$  представляется вектором (столбцом) матрицы  $W$ . Используя векторную модель, степень подобия  $coc(i,j)$  определим так:

$$coc(i,j) = \frac{\sum_k w_{ki}w_{kj}}{\sqrt{\sum_k w_{ki}^2} \sqrt{\sum_k w_{kj}^2}} \quad (1.5)$$

Матрица смежности взвешенного неориентированного графа, соответствующего СКЦЖ, обозначается  $W^{coc}$ . В качестве графа  $G^{coc} = (V^{coc}, E^{coc})$  рассматривается его максимальная связная компонента  $|V^{coc}| = 1582$  (остальные 147 вершин являются одиночными),  $|E^{coc}| = 820982$  ребер.

**2. Коэффициенты кластеризации.** Для вершины  $i$  графа  $G$  коэффициент кластеризации отражает вероятность того, что две ее соседние вершины  $G$  тоже являются соседями. Это локальное свойство вершины получило название „локальный коэффициент кластеризации“. Заметим, что термин „кластеризация“ не имеет здесь общепринятого значения. Определение коэффициента для невзвешенных графов без кратных ребер и петель приведено в работе [8]. Параметр характеризует степень связности графа.

2.1. *Локальный коэффициент кластеризации*  $CC(i)$  узла  $i$  определяется как отношение числа существующих ребер между соседями узла по отношению к максимально возможному числу таких ребер. Пусть  $N_i = \{j : (i,j) \in E \vee (j,i) \in E\}$  — множество соседей вершины  $i$ , а  $k_i = |N_i|$  — их число. Обозначим  $E_i$  число ребер между соседями. Для неориентированного графа максимально возможное число таких ребер равно  $k_i(k_i - 1)/2$ , а параметр  $CC(i)$  определяется так:

$$CC(i) = \frac{2E_i}{k_i(k_i - 1)} = \frac{|\{(j,k) \in E : j, k \in N_i\}|}{k_i(k_i - 1)} = \frac{\sum_{j,h} a_{ij}a_{ih}a_{jh}}{k_i(k_i - 1)}. \quad (2.1)$$

Для орграфа максимально возможное число ребер между соседями равно  $k_i(k_i - 1)$ , поэтому параметр  $CC(i)$  определяется так:

$$CC(i) = \frac{E_i}{k_i(k_i - 1)}. \quad (2.2)$$

*Локальный коэффициент кластеризации*  $CC(G)$  графа  $G$  определяется как среднее значение коэффициентов  $CC(i)$ :

$$CC(G) = \frac{1}{n} \sum_i CC(i). \quad (2.3)$$

В работе [8] коэффициент  $CC(G)$  рассматривается как один из параметров модели „малого мира“, которая характеризуется малым „средним расстоянием“  $L_{avg}(G)$  [2] между узлами и большим значением  $CC(G)$ , кроме того, при росте числа вершин параметр  $L_{avg}(G)$  растет медленно (логарифмическая зависимость), а  $CC(G)$  — быстро.

2.2. Альтернативный подход к определению связанности соседей представлен в работе [9], где введено отношение *транзитивности* между вершинами графа. Рассмотрим случай неориентированных графов. Отношение  $T$  обладает свойством транзитивности, если для любых  $i, j, k$  из того, что  $iTj$  и  $jTk$ , следует  $iTk$ . Для вершин это обозначает, что пути длины два замкнуты. Для примера укажем, как это свойство выполняется для всех путей длины два. Такие пути можно представить с помощью матрицы  $A^2$ , где  $A$  — матрица смежности  $G$ . Путь из  $i$  в  $j$  замкнут, если имеет место:  $A_{ij}^2 \geq 1$  &  $A_{ij} = 1$ .

В работе [10] *относительная транзитивность* графа выражается через треугольники и тройки вершин. *Треугольник*  $\Delta = \{V_\Delta, E_\Delta\}$  — это полный подграф графа  $G$ , множество вершин которого равно трем. *Тройка* — это подграф графа  $G$ , множество вершин которого равно трем, а множество ребер — двум. Тройка называется *тройкой с вершиной  $i$* , если  $i$  инцидентна обоим ребрам тройки. Число троек с вершиной  $i$ :

$$\tau(i) = C_2^{\deg(i)} = \frac{\deg(i)^2 - \deg(i)}{2}.$$

Число всех троек в графе  $G$  определяется суммой  $\tau(G) = \sum_i \tau(i)$  и совпадает с числом путей длины два. Обозначим через  $\lambda(G)$  число треугольников графа  $G$ . Определим  $\lambda(i) = \{|\Delta| : i \in V_\Delta\}$  — число треугольников, в которых  $i$  является одной из вершин. Заметим, что  $\lambda(G) = 1/3 \sum_i \lambda(i)$ . В этих терминах (2.1) выглядит как  $CC(i) = \frac{\lambda(i)}{\tau(i)}$ . Пусть

$$V' = \{i : \deg(i) \geq 2\}, \text{ тогда } CC(G) = \frac{1}{|V'|} \sum_{i \in V'} CC(i).$$

Глобальный коэффициент кластеризации графа (*транзитивность*), явно определенный в работе [11], задается равенством:

$$T(G) = \frac{\text{число треугольников}}{\text{число троек}} = \frac{3\lambda(G)}{\tau(G)}. \quad (2.4)$$

Отсюда  $0 \leq T(G) \leq 1$ . Заметим, что число треугольников можно вычислить как сумму диагональных элементов матрицы  $A^3$ , а по матрице  $A^2$  можно определить число троек.

Равенство, демонстрирующее разницу между локальным и глобальным коэффициентами кластеризации, приведено в работе [12]:

$$T(G) = \frac{\sum_{i \in V'} \tau(i) CC(i)}{\sum_{i \in V'} \tau(i)}. \quad (2.5)$$

Отсюда, глобальный и локальный коэффициенты будут равны, если все вершины имеют одинаковую степень или все локальные кластерные коэффициенты равны. Локальный коэффициент оценивает, насколько связны соседи для каждой вершины, а для глобального коэффициента все тройки эквивалентны.

Понятие локального коэффициента кластеризации расширено для случая взвешенной сети в работе [13]. Для взвешенных графов в формуле участвует *сила связей*  $s_i$  вершины  $i$ , которая для неориентированного графа определяется с помощью матрицы  $W$  как:

$$s_i = \sum_{j=1}^n w_{ij}. \quad (2.6)$$

Соответственно, *взвешенная степень вершины* неориентированного графа на основе „силы связей“ определяется из выражения  $\deg^w(i) = \frac{s_i}{\deg(i)}$ .

*Взвешенный локальный коэффициент кластеризации* узла  $i$  (неориентированный случай) определяется так:

$$CC^w(i) = \frac{1}{s_i(k_i - 1)} \sum_{j,h} \frac{w_{ij} + w_{ih}}{2} a_{ij} a_{ih} a_{jh}. \quad (2.7)$$

Этот коэффициент определяет не только число треугольников, одной из вершин которых является  $i$ , но и суммарный вес относительно силы взаимодействия. Нормализующий

элемент  $s_i(k_i - 1)$  вычисляет вес всех инцидентных ребер, умноженный на максимальное число треугольников с вершиной  $i$ , т. е.  $0 \leq CC^w(i) \leq 1$ .

Взвешенный локальный коэффициент кластеризации графа определяется следующим образом:

$$CC^w(G) = \frac{1}{n} \sum_i CC^w(i). \quad (2.8)$$

Отметим, что если  $\forall i, j w_{ij} = 1$ , то результаты (2.3) и (2.8) совпадают. Как замечено в работе [13], в реальных сетях, если имеет место неравенство  $CC^w(G) > CC(G)$ , то наиболее вероятно, что треугольники образуются ребрами с большими весами, если  $CC(G) > CC^w(G)$ , то топологическая кластеризация свойственна ребрам с маленькими весами. Другие подходы к определению локального коэффициента кластеризации для случая взвешенной сети можно найти в работах [14–16].

В работе [17] определен обобщенный глобальный коэффициент кластеризации как:

$$T_w = \frac{\text{общий вес треугольников}}{\text{общий вес троек}}.$$

Вес тройки определяется в зависимости от особенностей сети, так для сети, в которой вес ребра определяется в терминах „стоимости“ или времени, это может быть, например, максимум весов ребер, арифметическое или геометрическое среднее. Для сети, где вес определяет величину потока, скорее, это должен быть минимум. Следуя [13], при определении веса треугольника вес замыкающего ребра не учитывается. В любом из этих подходов для невзвешенного графа вес тройки будет равен единице и имеет место равенство  $T_w = T$ .

### 2.3. Результаты измерения коэффициентов кластеризации.

2.3.1. Локальный коэффициент кластеризации СЦЖ. Для вычисления взвешенного локального коэффициента кластеризации  $CC^w$  орграф  $G$  был преобразован в неориентированный  $G_u$  следующим образом. Если между вершинами  $i$  и  $j$  имеется только одна взвешенная дуга  $(i, j)$  или  $(j, i)$ , то дуга преобразуется в ребро  $(i, j)$ , которому присваивается вес, равный весу дуги; если между вершинами  $i$  и  $j$  имеются две дуги  $(i, j)$  и  $(j, i)$ , то обе они преобразуются в одно ребро  $(i, j)$ , которому присваивается вес  $w_{ij} + w_{ji}$ , равный сумме весов дуг. Значение  $CC^w(G_u) = 0,64170$  получено с помощью пакета *igraph* (согласно (2.7), (2.8)), сложность вычислений оценивается как  $\mathcal{O}(|V| \text{mean}(\text{deg}(\cdot))^2)$ . Отметим, что 2,9 % вершин  $G_u$  имеют менее двух соседей, для них значение коэффициента кластеризации считается неопределенным; у 1,7 % вершин соседние вершины не имеют между собой связей, для них  $CC^w(i) = 0$ . Заметим, что если веса ребер нормализовать, например, путем

деления веса ребра на средний вес ребра в графе  $w_{ij} = \frac{w_{ij}}{\sum_{ij} w_{ij}/|E|}$ , значение коэффициента

изменится,  $CC^w(G_u) = 0,48463$ .

Для вычисления коэффициента кластеризации невзвешенного графа, согласно (2.3), орграф  $G_u$  был преобразован ( $G_u^{runw}$ ), в результате: независимо от того, имеется между вершинами  $i$  и  $j$  одна взвешенная дуга или две, остается одно невзвешенное ребро. Значение параметра  $CC(G_u^{runw}) = 0,54594$ .

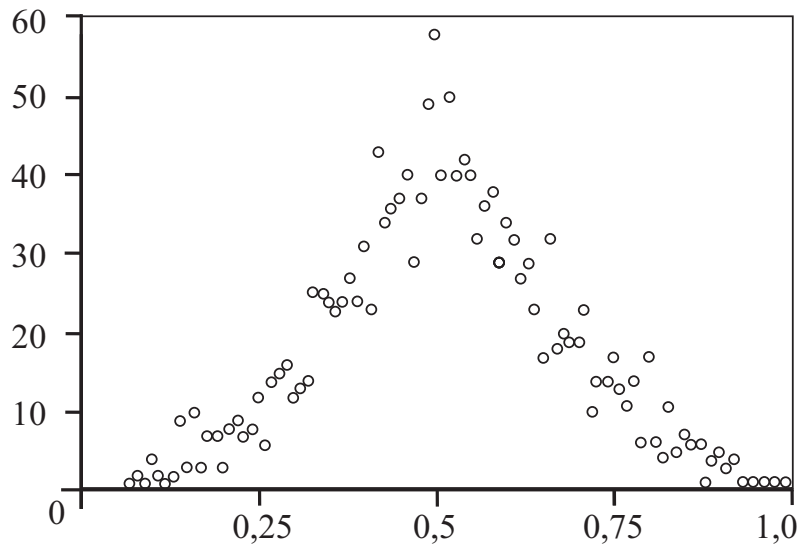


Рис. 2.1. Распределение  $CC^w(i)$  графа  $G$ . На оси абсцисс указаны абсолютные значения коэффициента; на оси ординат — число вершин, для которых  $x \leq CC^w(i) \leq x + 0,01$

Для уточнения зависимости между взвешенными степенями вершин  $deg^w(i) = \frac{\sum_{j=1}^n w_{ij}}{deg(i)}$  и взвешенными локальными коэффициентами кластеризации вычислен коэффициент ранговой корреляции Спирмена, его значение  $r = -0,352$  говорит о том, что, несмотря на выполнение неравенства  $CC^w(G_u) > CC(G_u^{unw})$ , для вершин с небольшой взвешенной степенью более характерно наличие взаимосвязанных соседей. Распределение взвешенных коэффициентов кластеризации представлено на рис. 2.1, на котором не представлены вершины с нулевым коэффициентом и вершины с максимальным значением, равным единице (48 вершин).

2.3.2. *Локальный коэффициент кластеризации СБСЖ.* Вычислены коэффициенты кластеризации для взвешенного и невзвешенного вариантов:  $CC^w(G^{bib}) = 0,80156$  (2.7, 2.8);  $CC^{unw}(G^{bib}) = 0,77971$  (2.3). Отметим, что 18,5% вершин  $G^{bib}$  имеют менее двух соседей, для них значение коэффициента кластеризации считается неопределенным. Разница между коэффициентами  $CC^w$  и  $CC^{unw}$  невелика. Коэффициент ранговой корреляции Спирмена, оценивающий зависимость между взвешенными степенями вершин и взвешенными локальными коэффициентами кластеризации, имеет значение  $r = -0,322$ .

2.3.3. *Локальный коэффициент кластеризации СКЦЖ.* Коэффициенты кластеризации для взвешенного и невзвешенного вариантов графа  $G^{coc}$ :  $CC^w(G^{coc}) = 0,82860$ ;  $CC^{unw}(G^{coc}) = 0,80588$ . При этом 9,3% вершин  $G^{coc}$  имеют менее двух соседей, для них значение коэффициента кластеризации считается неопределенным. Разница между коэффициентами для взвешенного и невзвешенного вариантов невелика. Коэффициент ранговой корреляции Спирмена, оценивающий зависимость между взвешенными степенями вершин и взвешенными локальными коэффициентами кластеризации, имеет значение  $r = -0,266$ .

**3. Кластерный анализ.** Широко используемый в прикладных исследованиях метод кластерного анализа (КА, или кластеризация) состоит в выявлении модульной структуры некоторого непустого множества объектов. Основой КА является группирование объектов на основании подобия их параметров. В нашем случае объектами являются научные жур-

налы, размещенные в одной БД. Кластеризация  $C = \{C_1, \dots, C_k\}$  графа  $G = (V, E)$  — это разбиение множества вершин  $V$  на желательные непересекающиеся непустые подмножества (кластеры)  $C_i$ . Разбиение множества вершин индуцирует разделение множества ребер. Обозначим  $E(C_i, C_j)$  множество ребер, ориентированных из  $C_i$  в  $C_j$ ;  $E(C_i, C_i)$  (сокращенно  $E(C_i)$ ), — множество ребер, связывающих вершины кластера  $C_i$ . Тогда  $E(C) := \cup_{i=1}^k E(C_i)$  является множеством внутрикластерных ребер, а  $E \setminus E(C)$  — множеством внекластерных ребер. Кластер будем идентифицировать с подграфом графа  $G$ , т. е.  $G[C_i] := (C_i, E(C_i))$ . Кластеризация называется *тривиальной*, если  $k = 1$ ; *одиночной*, если  $k = n$ ; *разрезом*, если  $k = 2$ .

Поскольку определение кластера не формализовано, существует ряд моделей кластеризации, отличающихся по двум основным признакам: что понимается под подобием и какие параметры рассматриваются. КА, базирующаяся на простой парадигме внутрикластерной плотности против внекластерной плотности, фокусируется на несвязанных между собой кликах как на идеальной ситуации. Применяемая техника кластеризации наиболее изучена и заключается либо в максимизации внутрикластерной плотности ребер, либо в минимизации внекластерной плотности. Предполагается, что тесно связанные сообщества с большей вероятностью имеют и другие общие свойства.

Кластеры, строящиеся на отношении подобия, отличном от плотности, учитывающем свойства самих узлов, например сходство их относительных позиций в сети, выявляют структуры с интересными свойствами коннективности. Для сетей цитирования это могут быть кластеры, строящиеся на отношении коцитирования (структурные свойства вершин кластера: являются конечными вершинами ребер, ведущих из одной и той же начальной вершины) или библиографического сочетания (структурные свойства вершин кластера: являются начальными вершинами ребер, ведущих в одну и ту же конечную вершину). К одной и той же сети могут быть применимы обе модели кластеризации. Обзор концепций, методов и алгоритмов, применяемых для кластеризации графов, можно найти в работах [18, 19].

Для проверки качества деления на кластеры в предположении, что истинные группы не известны, разрабатываются индексы качества кластеризации. В этом вопросе консенсус также не достигнут. Определения индексов, построенных на соотношении внутри и вне кластерных связей, можно найти, например, в работах [20, 21]. Распространенным способом проверки является вычисление параметра „модульность“, определенного в работе [22]. Определение строится в предположении, что структура графа, содержащего некие сообщества вершин, как правило, будет отклоняться от структуры случайного графа. Оценивается, насколько доля ребер между вершинами одного типа (т. е. попавших в один кластер) отличается от ожидаемой доли таких ребер в том случае, если ребра располагаются случайным образом, независимо от типа вершин. Для неориентированного графа параметр (нормализованная) *модульность* вычисляется по формуле:

$$Q^u = \frac{1}{2m} \sum_{i,j} \left( a_{ij} - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j), \quad (3.1)$$

где  $k_i$  — степень вершины  $i$ ,  $\delta(C_i, C_j) = 1$ , если  $C_i = C_j$ , и  $\delta(C_i, C_j) = 0$  в противном случае. Формула соответствует „модели предпочтительного присоединения“, в общем виде вместо  $\frac{k_i k_j}{2m}$  используется вероятность присоединения вершины  $i$  к  $j$ .

В работе [23] определение модульности расширено для орграфа:

$$Q^d = \frac{1}{m} \sum_{i,j} \left( a_{ij} - \frac{k_i^{\text{out}} k_j^{\text{in}}}{m} \right) \delta(C_i, C_j), \quad (3.2)$$

где  $k_i^{\text{out}}$  — исходящая степень вершины  $i$ ,  $k_j^{\text{in}}$  — входящая степень вершины  $j$ .

Для взвешенных графов с положительными весами в работе [24] предложено пользоваться формулами (3.1), (3.2), где  $A$  — взвешенная матрица смежности. Если используется отдельная матрица весов  $W$  и  $w_i = \sum_j w_{ij}$ ,  $w = \sum_{i,j} w_{ij}$ , то формула (3.1) будет иметь вид:

$$Q^u = \frac{1}{2w} \sum_{i,j} \left( w_{ij} - \frac{w_i w_j}{2w} \right) \delta(C_i, C_j). \quad (3.3)$$

Подход к определению модульности при наличии отрицательных весов представлен в работе [25].

Для сравнения результатов кластеризации одного и того же множества данных различными алгоритмами разработан ряд критериев. В нашем случае используются два индекса (меры) согласованности. Индекс  $NMI$ , представленный в работе [26], определяет степень согласованности двух делений на кластеры на основе понятия взаимной информации, используемой в теории вероятностей. Пусть имеется  $n$  объектов и два результата разделения на кластеры:  $A$  с кластерами  $C_1^A, C_2^A, \dots, C_k^A$  и  $B$ , с кластерами  $C_1^B, C_2^B, \dots, C_l^B$ . Информационная энтропия разделения на кластеры  $A$  определяется как

$$H(A) = - \sum_{i=1}^k \frac{n_i^A}{n} \log \left( \frac{n_i^A}{n} \right),$$

где  $n_i^A$  — число элементов в кластере  $C_i^A$ . Рассмотрим матрицу соответствия  $N^{AB}$  размерности  $k \times l$ , номера строк которой соответствуют номерам кластеров разделения  $A$ , столбцов — номерам кластеров разделения  $B$ , элемент  $N_{i,j}^{AB} = |C_i^A \cap C_j^B|$  — число объектов, общих для кластеров  $C_i^A$  и  $C_j^B$ . Индекс  $NMI(A, B)$  определяется равенством

$$NMI(A, B) = \frac{-2 \sum_{i=1}^k \sum_{j=1}^l N_{ij}^{AB} \log \left( \frac{N_{ij}^{AB} \times n}{N_i^A N_j^B} \right)}{\sum_{i=1}^k N_i^A \log \left( \frac{N_i^A}{n} \right) + \sum_{j=1}^l N_j^B \log \left( \frac{N_j^B}{n} \right)}, \quad (3.4)$$

где  $N_i^A$  — сумма по строке, а  $N_j^B$  — сумма по столбцу матрицы  $N^{AB}$ .

Индекс  $RAND(R)$ , представленный в работе [27], определяет долю согласованных пар кластеризуемых объектов по отношению ко всем парам. Пусть  $a$  — число пар, находящихся в одном кластере при делении  $A$  и делении  $B$ ;  $b$  — число пар, находящихся в разных кластерах при делении  $A$  и при делении  $B$ ;  $c$  — число пар, находящихся в одном кластере при делении  $A$  и в разных при делении  $B$ ;  $d$  — число пар, находящихся в одном кластере при делении  $B$  и в разных при делении  $A$ . Тогда

$$RAND(A, B) = \frac{a + b}{a + b + c + d} = \frac{a + b}{C_n^2}.$$

В терминах матрицы  $N^{AB}$ :

$$RAND(A,B) = \frac{C_n^2 + 2 \sum_{i=1}^k \sum_{j=1}^l C_2^{N_{ij}^{AB}} - \left[ \sum_{i=1}^k C_2^{N_i^A} + \sum_{j=1}^l C_2^{N_j^B} \right]}{C_n^2}. \quad (3.5)$$

3.1. *Алгоритмы кластеризации.* Следует заметить, что большинство определений и алгоритмов выявления сообществ предназначены для неориентированных невзвешенных графов. В нашем случае вес ребра является показателем силы связей узлов, которую следует учитывать. Рассмотрим четыре алгоритма выявления сообществ для взвешенных графов.

3.1.1. Иерархический алгоритм *BTW*, предложенный в работе [28], базируется на предположении, что меры центральности сетевых акторов могут использоваться для классификации узлов сети. Предлагаемый алгоритм основан на параметре „реберная центральность по посредничеству“ [29]. Параметр учитывает долю кратчайших путей между парой вершин, проходящих через данное ребро:

$$C_B(e) = \sum_{i \neq j \in V} \frac{\sigma_{ij}(e)}{\sigma_{ij}}, \quad (3.6)$$

где  $\sigma_{ij}$  — число кратчайших путей от вершины  $i$  до вершины  $j$  графа, а  $\sigma_{ij}(e)$  — число кратчайших путей от  $i$  до  $j$ , проходящих через ребро  $e$ . Предполагается, что ребра между сообществами имеют большее значение параметра. Основной цикл алгоритма выглядит так: 1) для всех ребер  $e$  вычисляется (3.6); 2) удаляется ребро с наибольшим значением  $C_B(e)$ ; 3) вычисляется значение модульности (3.2)–(3.3); 4) цикл повторяется. Таким образом, исполняется иерархическая процедура разделения. Лучшим считается уровень, соответствующий наибольшему значению модульности. Алгоритм используется как для неориентированных, так и для орграфов, сложность алгоритма оценивается как  $\mathcal{O}(|V||E|^2)$ .

3.1.2. В работе [30] рассматривается влияние структуры сети на сетевые потоки. Метод кластеризации использует информационную стоимость описания свободного блуждания по сети (или движения потоков) при различных делениях акторов на сообщества и связан с принципом минимального описания, формализованным в работе [31] и заключающимся в том, что лучшая гипотеза по поводу структуры данных та, которая ведет к большей компрессии данных. Оценивается нижняя граница кода  $L(M)$ , соответствующего разделению  $M = \{M_1, M_2, \dots, M_k\}$  имеющих  $n$  вершин на  $k$  кластеров, который учитывает энтропию свободного блуждания внутри и вне кластера и базируется на теореме Шеннона [32] о кодировании источника, определяющей нижнюю оценку кодирования  $n$  состояний случайной переменной через ее энтропию. Соответствующий базовый алгоритм *IMP* для неориентированных графов [30] состоит из двух этапов. Первоначально каждая вершина является кластером. На первом этапе случайным образом вершина объединяется с различными соседями в целях достижения наибольшего уменьшения  $L(M)$ ; на втором сеть перестраивается так, что каждый кластер становится вершиной. Процесс повторяется до тех пор, пока  $L(M)$  уменьшается. Для ориентированных сетей минимальное кодирование определяется на основе стационарного распределения вероятностей посещения узлов. При этом вводится вероятность  $\tau$  случайного перехода в другой узел, не следуя логике сети, гарантирующая уникальность такого распределения.

3.1.3. Подход, также основанный на процессе случайного блуждания, используется в работе [33]. Представленный алгоритм *WTR* выявляет плотные подграфы исходного графа. На начальном этапе считаем, что каждая вершина является кластером. Основной цикл

Таблица 3.1

Кластеризация орграфа  $G$  алгоритмом  $IMP$  [30]

$\#Cl$	$\#J$	Тематика
1	675	Финансовая экономика, обзорные журналы по экономике, математические и количественные методы
1	160	Администрирование и экономика бизнеса, маркетинг, бухгалтер
1	139	Евроэкономика (Румыния > 60 %)
1	131	Экономика сельского хозяйства и природных ресурсов
1	80	Транспортная экономика, математические и количественные методы
1	62	Здравоохранение, социальное обеспечение
1	60	Эконометрические и статистические методы
1	54	Региональная экономика
1	35	Энергетика
1	22	Математические и количественные методы
1	20	Чехия
1	19	Польша
1	14	Восточная Европа
2	13	Социальное обеспечение, немецкоязычные журналы
3	10	Администрирование и экономика бизнеса
2	9	Болгария, регулирование

Примечание. Здесь и далее:  $\#Cl$  — число кластеров,  $\#J$  — число журналов в кластере

алгоритма выглядит так: 1) вычисляются „расстояния“ между всеми соседними кластерами; 2) по критерию „расстояния“ выбираются два соседних кластера; 3) они объединяются в новый кластер и выполняется переход к шагу (1). Этот цикл повторяется  $(|V| - 1)$  раз. Для определения качества разбиения вычисляется модульность. Вычислительная сложность алгоритма равна  $\mathcal{O}(|E| |V|^2)$ . Для разреженных сетей вычислительная сложность оценивается как  $\mathcal{O}(|V|^2 \log(|V|))$ .

3.1.4. Иерархический алгоритм  $MLO$ , основанный на оптимизации модульности, предложен в работе [34]. Первоначально предполагается, что каждая вершина графа образует кластер. Алгоритм состоит из повторяющихся шагов, каждый шаг состоит из двух фаз.

Фаза 1. Для каждой вершины  $i$  рассматривается перемещение в кластер  $C$ , соответствующий ближайшему соседу  $j$ . Вычисляется изменение модульности при перемещении вершины  $i$  в кластер  $C$ . Проверяются все ближайшие соседи  $i$ , перемещение происходит в тот кластер, которому соответствует наибольшее увеличение модульности. Если улучшение невозможно,  $i$  остается в своем кластере. Рассматриваются все вершины до тех пор, пока возможно улучшение. Первая фаза заканчивается достижением локального максимума модульности.

Фаза 2. Построение нового графа, вершины которого соответствуют кластерам, полученным во время первой фазы. Ребра между вершинами двух кластеров заменяются на одно ребро, вес которого равен сумме их весов, ребра между вершинами одного кластера заменяются на петлю, вес которой равен сумме весов внутренних ребер.

Фазы 1–2 повторяются до тех пор, пока невозможны новые изменения. Число кластеров существенно уменьшается с каждым шагом. Сложность алгоритма в применении к разреженным графам составляет  $\mathcal{O}(|E|)$ .

### 3.2. Результаты кластеризации.

3.2.1. *Кластеризация СЦЖ.* Кластеризация орграфа  $G$  производилась с помощью алгоритмов *IMP* и *BTW*. В результате кластеризации  $G$  алгоритмом *IMP* было получено 20 кластеров размером 9 и более вершин (89,35 %). Остальные 97 кластеров имеют меньшее число вершин. Результаты проверки кластеризации алгоритмом *IMP* показала, что кластеры средних размеров выявляют сообщества журналов, публикующих, в основном, статьи, относящиеся к определенным тематикам или территориальным группам. Наблюдаемая тенденция: у кластеров с числом вершин, большим 22, число вершин, у которых есть исходящие дуги, ведущие из кластера (*шлюзы*), меньше, чем число вершин, к которым есть входящие извне дуги; такое же распределение у суммарного веса соответствующих дуг. У небольших кластеров наоборот: суммарный вес выходящих из кластера дуг больше, чем входящих, т. е. небольшие кластеры представляют собой более замкнутые тематические сообщества. Результаты приведены в табл. 3.1.

Алгоритм *BTW* основан на подсчете кратчайших путей, поэтому вместо весов дуг  $w_{ij}$  использовались обратные значения  $1/w_{ij}$  (см. [35]). В результате кластеризации графа  $G$  получен один кластер размером 1242 вершины, что составляет 71,83 % вершин. Остальные 483 кластера имеют 4 и менее вершин, большинство из них одновершинные. Индексы согласованности кластеризации  $G$  алгоритмами *IMP* и *BTW*:  $NMI = 0,23$ ;  $Rand = 0,44$ . Алгоритм *BTW* применен также к орграфу  $G$ , где веса ребер преобразованы согласно [36]:  $w_{ij} = w_{\max} + 1 - w_{ij}$ , где  $w_{\max}$  – максимальный вес дуг графа  $G$ . Результаты кластеризации изменились: наибольший кластер содержит 1522 вершины (80,02 %), остальные содержат 3 вершины и менее. Более того, подобный результат получился и при применении алгоритма *BTW* к невзвешенному орграфу  $G^{unw}$ : большинство вершин попали в один большой кластер (87,85 %). Можно предполагать, что в нашем случае алгоритм нечувствителен к масштабу весов дуг.

Для неориентированного варианта графа цитирования  $G_u$  (см. 4.1.1) рассмотрены четыре алгоритма кластеризации: *BTW*, *IMP*, *WTR* и *MLO*. При кластеризации  $G_u$  алгоритмом *BTW* вновь получилась практически тривиальная кластеризация: 86 % вершин попали в один кластер, наибольший процент остальных — в одиночные кластеры. Распределение размеров кластеров, полученных в результате работы остальных алгоритмов, представлено в табл. 3.2.

Индексы согласованности результатов кластеризации  $G_u$  выглядят так:  $NMI = 0,52$ ,  $RAND = 0,84$  для *WTR* и *MLO*;  $NMI = 0,62$ ,  $RAND = 0,86$  для *IMP* и *MLO*;  $NMI = 0,58$ ,  $RAND = 0,86$  для *IMP* и *WTR*. Согласованность алгоритма *IMP* в применении к графам  $G$  и  $G_u$ :  $NMI = 0,54$ ;  $RAND = 0,82$ , — т. е. сравнима с кластеризацией  $G_u$  с помощью остальных алгоритмов. Несмотря на различие распределения размеров кластеров, тематические области сходны. Области определялись на основании названий журналов, тематика сопоставляется согласно большинству в процентном отношении. Следует отметить, что для журналов издательства *Elsevier* (153 журнала) в базе данных указаны коды тематик согласно классификационной системе *Jel* [37]. Анализ показал, что коды журналов издательства *Elsevier* соответствуют установленным тематикам кластеров, в которые они попали. При этом установленная тематика в большинстве случаев относится к основным тематикам (*general categories*) по классификации *Jel*. Тематические группы, выявленные алгоритмом *IMP* на графе  $G_u$ , представлены в табл. 3.3.

3.2.2. *Кластеризация  $G^{bib}$ .* Алгоритм *IMP* объединил все 1432 вершины в один кластер, т. е. не установил дополнительного подобия уже подобных согласно отношению  $R^{bib}$  вер-

Таблица 3.2

Распределение размеров кластеров  
при кластеризации графа  $G_u$

IMP [30]		WTR [33], $Q^u = 0,42$		MLO [34], $Q^u = 0,43$	
#Cl	#J	#Cl	#J	#Cl	#J
1	737	1	657	1	585
1	186	1	284	1	395
1	158	1	268	1	226
1	136	1	221	1	178
1	95	1	75	1	104
1	65	1	62	1	100
1	60	1	44	1	85
1	59	2	8	1	54
1	48	1	1	1	2
1	38				
1	35				
1	32				
1	15				
1	14				
1	12				
1	9				
3	4				
2	3				
6	2				

шин. Размеры кластеров, полученных с применением алгоритмов *WTR* и *MLO*, приведены в табл. 3.4.

Согласованность *WTR* и *MLO*:  $NMI = 0,79$ ;  $RAND = 0,84$ . Если сравнить кластеры *WTR* (30 вершин) и *MLO* (70 вершин), то в первом кластере 67% журналов относятся к статистическим исследованиям, а во втором 46% таких журналов, причем все 30 журналов первого кластера содержатся во втором, 24% журналов относятся к эконометрии.

3.2.3. *Кластеризация  $G^{coc}$* . Использованы алгоритмы, приведенные в п. 3.2.2. Алгоритмом *IMP* получена тривиальная кластеризация. Размеры кластеров, полученных с применением алгоритмов *WTR* и *MLO*, приведены в табл. 3.5.

Кластеризация с помощью алгоритмов *WTR* и *MLO* получена при значении модульности  $Q = 0,16$  и  $Q = 0,21$  соответственно; согласованность:  $NMI = 0,64$ ;  $RAND = 0,82$ .

**Заключение.** Проанализированы библиометрические сети, акторами которых являются журналы, проиндексированные в базе данных RePEc. Они представлены плотными графами, имеющими большие значения коэффициентов кластеризации, независимо от того, учитываются ли веса ребер. Выявление сообществ с помощью алгоритма *IMP* в применении к взвешенному орграфу цитирования позволило выявить сообщества журналов, тесно связанных по тематическому или территориальному признаку. Алгоритм *BTW* не дал явных результатов ни в применении к орграфу цитирования, ни к его неориентированному представлению. В то же время результаты алгоритмов *IMP*, *WTR* и *MLO* в применении к неориентированному взвешенному графу имеют достаточный уровень со-

Таблица 3.3

Тематика кластеров при кластеризации  $G_u$  алгоритмом  $IMP$ 

#Cl	#J	Тематика
1	737	Финансовая экономика, математические и количественные методы
1	186	Администрирование и экономика бизнеса, маркетинг, бухгалтер
1	158	Финансовая экономика
1	136	Экономика сельского хозяйства и природных ресурсов
1	95	Румыния
1	65	Эконометрические и статистические методы
1	60	Здравоохранение, социальное обеспечение
1	59	Исследование операций
1	48	Региональная экономика
1	38	Энергетика
1	35	Транспортная экономика, математические и количественные методы
1	32	Администрирование и экономика бизнеса, маркетинг, бухгалтер
1	15	Восточная Европа
1	14	Жилищная и имущественная политика
1	12	Не идентифицирован
1	9	Образование

Таблица 3.4

Кластеризация графа  $G^{bib}$ 

WTR [33], $Q^u = 0,13$		MLO [34], $Q^u = 0,15$	
#Cl	#J	#Cl	#J
1	718	1	541
1	369	1	514
1	288	1	307
1	30	1	70
1	25		
2	1		

Таблица 3.5

Кластеризация графа  $G^{coc}$ 

WTR [33], $Q^u = 0,13$		MLO [34], $Q^u = 0,15$	
#Cl	#J	#Cl	#J
1	517	1	410
1	350	1	391
1	255	1	369
1	235	1	315
1	32	1	97
1	30		
1	9		
9	< 9		

гласованности, особенно соответственно индексу  $RAND$ . Во всех трех случаях получился один большой кластер, содержащий более 33 % вершин; остальные вершины распределены по сообществам, причем алгоритм  $MLO$  выделил наибольшее число сообществ. Сравнение результатов кластеризации алгоритмом  $IMP$  в применении к орграфу цитирования и неориентированному графу показало, что, несмотря на различие в размерах кластеров и распределении вершин по кластерам, определены сходные тематики. Таким образом, неориентированный вариант графа вполне можно использовать для выявления тематик.

Размер кластеров, полученных в результате кластеризации  $G^{bib}$  и  $G^{coc}$ , слишком большой, чтобы соотнести их с тематическими группами, однако согласованность примененных алгоритмов достаточно высокая. Следует заметить, что кластеризация сетей библиографического сочетания и коцитирования, акторами которых являют-

ся публикации из рассмотренных здесь журналов [38], была получена с применением алгоритма *MLO* при больших значениях модульности и большей степени согласованности.

## Список литературы

1. REPEC. General principles. [Electron. resource]. <http://repec.org/>.
2. БРЕДИХИН С. В., ЛЯПУНОВ В. М., ЩЕРБАКОВА Н. Г. Структура сети цитирования научных журналов // Проблемы информатики. 2017. № 2. С. 38–52.
3. HARARY F. Graph Theory. Addison-Wesley, 1969.
4. KESSLER M. M. Bibliographic coupling between scientific papers // American Documentation. 1963. V. 14. P. 10–25.
5. SALTON G., MACGILL M. J. Introduction to modern information retrieval. N. Y.: McGraw-Hill, 1983.
6. SMALL H. Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents // J. of the American Society for Information Science. 1973. V. 24. P. 265–269.
7. MARSHAKOVA I. System of document connections based on references // Scientific and Technical Information Serial of VINITI. 1973. V. 6, N 2. P. 3–8.
8. WATTS D. J., STROGATZ S. H. Collective dynamics of „small-world“ networks // Nature. 1998. V. 393. P. 440–442.
9. WASSERMAN S., FAUST K. Social network analysis: Methods and applications. Cambridge (ENG), New York: Cambridge University Press, 1994.
10. BRANDES U. Network analysis. Berlin, Heidelberg, New York: Springer, 2005.
11. NEWMAN M. E. J., STROGATZ S. H., WATTS D. J. Random graph models of social networks // Proc. of the National Academy of Science of the USA. 2002. V. 99. P. 2566–2572.
12. BOLLOBAS B., RIORDAN O. M. Mathematical results on scale-free random graphs. Handbook of graphs and networks: From genome to Internet. Weinheim, FRG: Wiley-VCH Verlag GmbH & Co. KGaA, 2002. P. 1–34.
13. BARRAT A., BARTHELEMY M., PASTOR-SATORRAS R., VESPIGNANI A. The architecture of complex weighted networks // Proc. of the National Acad. of Sci. 2004. V. 101, iss. 11. P. 3747–3752.
14. LOPEZ-FERNANDEZ L., ROBLES G., GONZALEZ BARAHONA J. Applying social network analysis to the information in cvs repositories // Proc. of the 1st Intl. workshop on mining software repositories (MSR2004). USA: Springer, 2004. P. 101–105.
15. ONNELA J.-P., SARAMI J., KERTZ J., KASKI K. Intensity and coherence of motifs in weighted complex networks // Phys. Rev. 2005. E 71 065103.
16. ZHANG B., HORVATH S. A general framework for weighted gene co-expression network analysis // Statistical Applications in Genetics and Molecular Biology. 2005. V. 4., N 17.
17. OPSAL T., PANZARASA P. Clustering in weighted networks // Social networks. 2009. V. 31. P. 155–163.
18. FORTUNATO S. COMMUNITY DETECTION IN GRAPHS // Physics Reports. 2010. V. 486. P. 75–174.
19. MALLIAROS F. D., VAZIRGIANNIS M. Clustering and community detection in directed networks: A survey // Physics Reports. 2013. V. 533, iss. 4. P. 95–142.
20. KANNAN R., VAMPALA S., VETTA A. On clustering — good, bad and spectral // Foundations of Computer Science. 2000. P. 367–378.
21. VAN DONGEN S. M. Graph clustering by flow simulation // PhD thesis. University of Utrecht. 2000.
22. NEWMAN M. E. J., GIRVAN M. Finding and evaluating community structure in networks // Phys. Rev. 2004. E 69 (2) 026113.

23. ARENAS A., DUCH J., FERNÁNDEZ A., GÓMEZ S. Size reduction of complex networks preserving modularity // *New J. Phys.* 2007. V. 9, N. 6. P. 176–190.
24. NEWMAN M. E. J. Analysis of weighted networks // *Phys. Rev. E* 70, 056131. 2004.
25. GÓMEZ S., JENSEN P., ARENAS A. Analysis of community structure in networks of correlated data // *Phys. Rev. E* 80, 016114.
26. FRED A. L. N., JAIN A. K. Robust data clustering // *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, Minneapolis, USA, June 16–22, 2003.* P. 128–136.
27. RAND W. M. Objective criteria for the evaluation of clustering methods // *J. Amer. Statistical Association.* 1971. V. 66, N. 336. P. 846–850.
28. GIRVAN M., NEWMAN M. E. J. Community structure in social and biological networks // *Proc. Nat. Acad. Sci. USA.* 2002. V. 99. P. 7821–7826.
29. FREEMAN L. C. A set of measures of centrality based upon betweenness // *Sociometry.* 1977. V. 40. P. 35–41.
30. ROSVALL M., BERGSTROM C. T. Maps of random walks on complex networks reveal community structure // *Proc. Natl. Acad. Sci. USA.* 2008. V. 105, N 4. P. 1118–1123.
31. RISSANEN J. Modeling by short data description // *Automatica.* 1978. V. 14. P. 465–471.
32. SHANNON C.E. A mathematical theory of communications // *Bell System Tech. J.* 1948. V. 27. P. 379–423.
33. PONS P., LATAPY M. Computing communities in large networks using random walks // *J. of Graph Algorithms and Applications.* 2006. V. 10, N 2. P. 191–218.
34. BLONDEL V., GUILLAUME J., LAMBIOTTE J., LEFEBVRE E. Fast unfolding of communities in large networks // *J. Stat. Mech.* 2008, P10008.
35. NEWMAN M. E. J. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality // *Phys. Rev. E* 64, 016132. 2001.
36. BRANDES U. On variants of shortest-path betweenness centrality and their generic computation // *Soc. Networks.* 2008. V. 30. P. 136–145.
37. JEL CLASSIFICATION SYSTEM / *EconLit Subject Descriptors.* 2016. [Electron. resource]. <https://www.aeaweb.org/econlit/jelCodes.php?view=jel>.
38. БРЕДИХИН С. В., ЛЯПУНОВ В. М., ЩЕРБАКОВА Н. Г. Структура сети цитирования научных статей // *Пробл. информ.* 2016. № 3. С. 28–43.



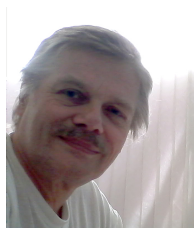
**Бредихин Сергей Всеволодович** — канд. техн. наук, ведущий научный сотрудник Ин-та вычислительной математики и математической геофизики СО РАН; e-mail: [bred@nsc.ru](mailto:bred@nsc.ru);

**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. В период 1988–2017 гг. руководил лабораторией ИВМ и МГ СО РАН. Исполнял обязанности технического директора проекта „Сеть Интернет Новосибирского научного центра“. Лауреат государственной премии РФ

по науке и технике 2012 г. Сфера научных интересов: анализ и измерение распределенных информационных сетей. Автор и соавтор более ста научных 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, and became an employee of Institute of Automation and Electrometry SB RAS. In 1983 he received PhD degree in Engineering Science. Since 1988–2017 he was the head of the laboratory of Computing Center (now ICM&MG) SB RAS. He was the technical manager of „Akademgorodok Internet Project“. He is the state prize winner in science

and engineering RF at 2012. Sphere of his scientific interests — analysis and measurement of the distributed information networks. He is the author and co-author of more than hundred scientific works and two monographs: „Metody bibliometrii i rynek elektronnoy nauchnoy periodiki“, „Analiz 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 г. 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 автоматки 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“. Научные 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.

*Дата поступления — 25.07.2017*