THE STRUCTURE OF THE CITATION NETWORK OF SCIENTIFIC PUBLICATIONS

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

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

Methods of measurement of the parameters characterizing a structure of the citation network of scientific publications are presented: average distance, density and transitivity. Values of these parameters are calculated based on the citation data extracted from the bibliographic DB *RePEc*. Clustering analysis of co-citation, bibliographic coupling and summary graphs corresponding to the main network component is made using two algorithms of community detection. The comparison of results was done by computing *NMI*. Analysing allowed to detect groups of articles, joint by common subject and to characterize them.

Key words: average distance, density, transitivity, clustering coefficient, communities, clustering algorithm, modularity, *NMI* measure.

References

1. Bredihin S. V., Lyapunov V. M., Sherbakova N. G., Yurgenson A. N. Parametry "central'nosti" uzlov seti citirovaniya nauchnyh statey // Problemy informatiki. 2016. № 1. S. 30–57.

2. Bredihin S. V., Lyapunov V. M., Sherbakova N. G. Parametry par uzlov seti citirovaniya nauchnyh statey // Problemy informatiki. 2016. № 2. S. 30–49.

3. General principles. [Electron. resource]. http://repec.org.

4. Milgram S. The small world problem // Psychol. Today. 1967. V. 2. P. 60–67.

5. Fortunato S. Community detection in graphs // Phys. Reports. 2010. V. 486. P. 75–174.

6. Watts D. J. Small worlds: The dynamics of networks between order and randomness. Princeton: Princeton University Press, 1999.

7. Wasserman S., Faust K. Social network analysis: Methods and applications. Cambridge: Cambridge University Press, 1994.

8. Broder A., Kumar R., et al. Graph structure in the web // 9th International World Wide Web conference, Amsterdam (Netherlands), 2000. V. 33. P. 309–320.

9. Faloutsos M., Faloutsos P., Faloutsos C. On power-law relationships of the internet topology // ACM conference on applications, technologies, architectures and protocols for computer communications, Cambridge (Engl.), 1999. P. 251–262. [Electron. resource]. http://www.cs.cmu.edu/~christos/publications/sigcomm99.pdf.

10. Watts D. J., Strogatz S. H. Collective dynamics of„small-world" networks // Nature. 1998. V. 393. P. 440–442.

11. Newman M. E. J. The structure and function of complex networks // SIAM Review. 2003. V. 45. P. 167–256.

12. Ebel H., Mielsch L. I., Bornholdt S. Scale-free topology of e-mail networks //Phys. Rev. E. 2002. V. 66, 035103.

13. Albert R., Barabasi A. L. Statistical mechanics of complex networks // Reviews of Modern Physics. 2002. V. 74. P. 47–97.

14. Burt R. S. The social structure of competition. Cambridge, MA: Harvard University Press, 1992.

15. Csárdi G., Nepusz T. The igraph software package for complex network research // InterJournal Complex Systems. 2006. 1695 P. [Electron. resource]. http://igraph.org/r/doc/.

16. Network analysis. Methodological Foundations. 2005. Springer, LNCS 3418.

17. Meilă M., Pentney W. Clustering by weighted cuts in directed graphs // Proc. of the 2007 SIAM International conference on data mining, 2007. Apr. 26–28. Minneapolis (USA). P. 135–144.

18. Marshakova I. V. Sistema svyazey mezhdu dokumentami, postroennaya na osnove ssylok: po dannym Science Citation Index // NTI, ser. 2. 1973. № 6. S. 3–8.

19. Small H. Co-citation in the scientific literature: A new measure of the relationship between two documents // J. Amer. Soc. Inform. Sci. 1973. V. 24, iss. 4. P. 265–269.

20. Kessler M. M. Bibliographic coupling between scientific papers // Amer. Documentation. 1963. V. 14, iss. 1. P. 10–25.

21. Satuluri V., Parthasarathy S. Symmetrizations for clustering directed graphs // Proc. 14th Internat. Conference on extending database technology, Uppsala (Sweden), March 21–25, 2011. P. 343–354. [Electron. resource]. http://dblp.uni- trier.de/db/conf/edbt/edbt2011.html.

22. Kleinberg J. M. Authoritative sources in a hyperlinked environment // J. of the ACM. 1999. V. 46, iss. 5. P. 604–632.

23. Zhou D., Schulkopf B., Hofmann T. Semi-supervised learning on directed graphs // Advances in Neural Information Processing Systems Conference, Dec. 5–8, 2005. Vancouver (Canada). P. 1633–1640.

24. Guimera R., Pardo M. S., Amaral L. A. N. Module identification in bipartite and directed networks // Phys. Rev. E 76 (3) 036102+. 2007.

25. Newman M.E.J., Girvan M. Finding and evaluating community structure in networks // Phys. Rev. 2004. E 69 (2) 026113.

26. Newman M. E. J. Fast algorithm for detecting community structure in networks // Phys. Rev. 2003. E 69 066133.

27. 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.

28. Leskovec J., Lang K. J., Dasgupta A., Mahoney M. W. Statistical properties of community structure in large social and information networks // Proc. of the 17th International Conference on World Wide Web, Beijing (China), April 21-25, 2008. P. 695–704.

29. Yang Y., Leskovec J. Overlapping community detection at scale: A nonnegative matrix factorization approach // Proc. of the 6th ACM International conference on web search and data mining, Rome (Italy). Feb. 6–8, 2013. P. 587–596.

30. Meila M. Comparing clusterings by the variation of information // Proc. of 16th annual conference on learning theory and 7th Kernel workshop, Washington (USA), Aug. 24–27, 2003. P. 173–187.

31. Fred A. L. N., Jain A. K. Robust data clustering // Proc. IEEE Computer Society conference on computer vision and pattern recognition, Minneapolis (USA), June 16–22, 2003. P. 128–136.

32. Girvan M., Newman M. E. J. Community structure in social and biological networks // Proc. Nat. Acad. Sci. USA. 2002. V. 99. P. 7821–7826.

33. Pons P., Latapy M. Computing communities in large networks using random walks // J. Graph Algorithms and Applications. 2006. V. 10, N 2. P. 191–218.

34. Chen J., Yuan B. Detecting functional modules in the yeast protein-protein interaction network // Bioinformatics. 2006. V. 22, iss. 18. P. 2283–2290.

35. Raghavan U. N., Albert R., Kumara S. Near linear time algorithm to detect community structures in large-scale networks // Phys. Rev. E 76, 036106. 2007.

36. Blondel VD., Guillaume J.L., Lambiotte R., Lefebvre E. Fast unfolding of community hierarchies in large networks // J. Stat. Mech. 2008. P10008.

СТРУКТУРА СЕТИ ЦИТИРОВАНИЯ НАУЧНЫХ СТАТЕЙ

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

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

УДК 001.12+303.2

Представлены методы измерения параметров, определяющих структуру сети цитирования научных статей: среднее расстояние, плотность и транзитивность. На основе данных о цитировании библиографической базы данных *RePEc* вычислены их значения. Для главной сетевой компоненты построены графы коцитирования, библиографического сочетания и выполнен их кластерный анализ с использованием двух алгоритмов. Произведено сравнение алгоритмов с помощью *NMI*. Результат кластеризации позволил выявить группы публикаций, объединенных общей тематикой, и охарактеризовать их.

Ключевые слова: среднее расстояние, плотность, кластерный коэффициент, сообщества, алгоритм кластеризации, модульность, мера *NMI*.

Введение. Продолжаем изучение сети цитирования научных статей (СЦС), начатое в работах [1, 2]. Данная работа посвящена анализу структуры СЦС, находящейся в состоянии постоянного роста за счет пополнения и регулярного индексирования информации о цитировании в библиографической базе данных (БД) *RePEc* [3]. Интерес к сетевым структурам, подобным СЦС, возник в связи с экспериментом С. Милгрэма по изучению распределения длин кратчайших путей между случайно выбранными вершинами связного графа [4]. В результате появилась гипотеза о сетевых объектах, имеющих достаточно короткий путь из любой вершины в любую другую. Такие объекты, как правило, имеют одну (главную) иерархически организованную связную компоненту Z значительного размера и характерные значения следующих параметров: достаточно малое значение диаметра графа, соответствующего Z; большое значение коэффициента кластеризации Z; распределение степеней вершин Z, отвечающее закону $x^{-\alpha}$. В дальнейшем они получили название„малый мир“.

В работе представлены методы измерения указанных параметров и вычислены их значения для СЦС, по которым можно судить, насколько изучаемый объект отвечает характеристикам„малого мира“. Выполнен кластерный анализ главной компоненты СЦС, определивший группы статей, объединенных общей тематикой. В конечном счете представленная методика позволяет выявить наличие сотрудничества и определить сетевые меры важности научных публикаций. Фундаментальный обзор по этой теме представлен в работе [5].

Напомним, СЦС представлена в виде орграфа $G = (V, E)$ с матрицей смежности $M = [m_{ij}], m_{ij} = 1$, если ребро $(j, i) \in E$ (статья j цитирует статью i) и $m_{ij} = 0$ в противном случае. На момент извлечения данных из БД орграф G имел одну главную слабо-связную компоненту A, состоящую из 131 684 вершин и 514 158 ребер, одну компоненту A_{16}, состоящую из 16 вершин и 31 ребра. Остальные компоненты имеют меньшие

Таблица 1.1

Параметры компонент A и A_{16}
(d — ориентированный, u — неориентированный графы)

A^u			A^d		A^u_{16}			A^d_{16}	
deg_{avg}	L_{avg} (1.1)	Δ^u (1.3)	L_{avg} (1.1)	Δ^d (1.2)	deg_{avg}	L_{avg} (1.1)	Δ^u (1.3)	L_{avg} (1.1)	Δ^d (1.2)
7,8	16,7	0,000059	19,7	0,000029	3,9	1,4	0,258333	2,0	0,129167

значения параметров и не рассматриваются. Количество ребер обозначаем m, количество вершин — n. Анализ главной компоненты выполнен с учетом ориентации ребер A^d и без учета — A^u. Аналогичные обозначения примем для компоненты A_{16}, которая используется для примеров. В зависимости от применяемого алгоритма значения ряда параметров приводятся в двух вариантах. Граф цитирования компоненты A_{16} и таблица связей меток вершин с библиографическими данными статей определены в работе [2].

1. Параметры„среднее расстояние“ и„плотность“. Согласно [6], *среднее расстояние* между вершинами графа G определяется как

$$L_{avg}(G) = \frac{1}{n(n-1)} \sum_{i \in V} \sum_{j \neq i \in V} d(i,j), \tag{1.1}$$

где $d(i,j)$ — расстояние от вершины i до вершины j; для орграфа $d(i,j) = 0$, если нет пути от i до j.

Плотность [7] орграфа G без кратных ребер и петель определяется как

$$\Delta^d(G) = \frac{|E|}{|V|(|V|-1)} = \frac{\sum\limits_j \sum\limits_i m_{ji}}{|V|(|V|-1)}, \tag{1.2}$$

для неориентированного графа

$$\Delta^u(G) = \frac{2|E|}{|V|(|V|-1)} = \frac{2\sum\limits_j \sum\limits_i m_{ji}}{|V|(|V|-1)}. \tag{1.3}$$

Разреженным считается граф, у которого $|E| << |V|^2$.

Для графов A^u и A^u_{16} вычислены параметры среднее расстояние L_{avg} и плотность Δ, а также среднее значение степени вершины deg_{avg}; значения приведены в табл. 1.1. Вывод: компоненты A^u и A^u_{16} представляют разреженные графы со сравнительно небольшим средним расстоянием между вершинами, что характерно для многих социальных сетей (всегда меньше 20, обычно меньше 10). Примером могут служить Всемирная паутина (*World Wide Web*) [8] и топология Интернет [9].

2. Параметр„кластерный коэффициент“. *Локальный кластерный коэффициент* C_i для неориентированного графа определяется в работе [10] как отношение количества существующих ребер E_i между соседями узла i к максимально возможному числу таких ребер. Максимум связей между соседями выражается отношением $k_i(k_i-1)/2$, где k_i — количество соседей вершины i (степень вершины). Таким образом, локальный кластерный коэффициент для вершины i неориентированного графа определяется как

$$C_i = \frac{2E_i}{k_i(k_i - 1)}. \tag{2.1}$$

В большинстве случаев ориентация ребер при вычислении кластерного коэффициента для орграфа игнорируется. *Среднее значение коэффициента* C_{avg} для вершин определяет степень кластеризации графа:

$$C_{avg} = \frac{1}{n}\sum_i C_i. \tag{2.2}$$

В работе [10] представлена модель динамических систем, которым свойственен эффект„малого мира“. Они характеризуются малым значением L_{avg} и большим значением C_{avg}; при росте числа вершин L_{avg} растет медленно, а C_{avg} быстро.

Альтернативное определение кластерного коэффициента графа G дано в книге [7]. Это — доля замкнутых путей длины два в сети, т. е. подсчитываются все пути длины два, и рассматривается, какая часть из них является замкнутыми:

$$C(G) = \frac{(\text{количество замкнутых путей длины два})}{(\text{количество всех путей длины два})}, \tag{2.3}$$

Заметим, что все пути длины два можно представить с помощью матрицы M^2. Путь из i в j замкнут, если при выполнении неравенства $m_{ij}^2 \geq 1$ имеет место $m_{ij} = 1$ (или $m_{ji} = 1$, когда отношение несимметрично). Значение кластерного коэффициента графа варьируется от нуля до единицы. Если $C = 1$, то имеет место полная транзитивность, т. е. граф представляет собой клику (все вершины связаны со всеми). Равенство $C = 0$ означает, что в графе нет замкнутых путей длины два; такая структура характерна, например, для дерева. Анализ социальных сетей показывает, что многие из них имеют достаточно высокое значение кластерного коэффициента (см. [11–13]). Заметим, что здесь термин„кластеризация“ не имеет общепринятого значения, а является синонимом термина„транзитивность“, так как исследуется, насколько часто из того, что узел i связан отношением с узлом j, а j связан с k, следует, что i связан с k.

Локальный кластерный коэффициент C_i может рассматриваться как мера центральности вершины графа (меньшие значения соответствуют более сильным акторам). C_i подобен центральности по посредничеству [1], отражающей интенсивность потока, циркулирующего между всеми парами вершин графа. С помощью C_i можно„измерять“ информационный поток, которым обмениваются непосредственные соседи актора. Корреляция коэффициентов центральность по посредничеству и локального кластерного коэффициента отмечена в работе [14]. Для многих сетей эмпирически выявлена обратная зависимость локального кластерного коэффициента от степени вершин, т. е. вершины с большей степенью в среднем имеют меньший локальный кластерный коэффициент.

Локальный кластерный коэффициент СЦС определяет, как часто пары статей, находящихся в отношении цитирования с рассматриваемой статьей, находятся в отношении цитирования друг с другом. Приведем результаты вычисления значений C_i (2.1) и C_{avg} (2.2) для графа A^u. Для этого воспользуемся пакетом *igraph* [15]. Гистограмма значений C_i представлена на рис. 2.1. Компонента A^u содержит 10,9 % вершин, имеющих меньше двух соседей, они не включены в гистограмму. Также исключены 27,5 % вершин, имеющих соседей, не связанных между собой. По оси абсцисс отложены значения C_i, а по оси ординат — соответствующее число вершин N. Из множества вершин, имеющих больше

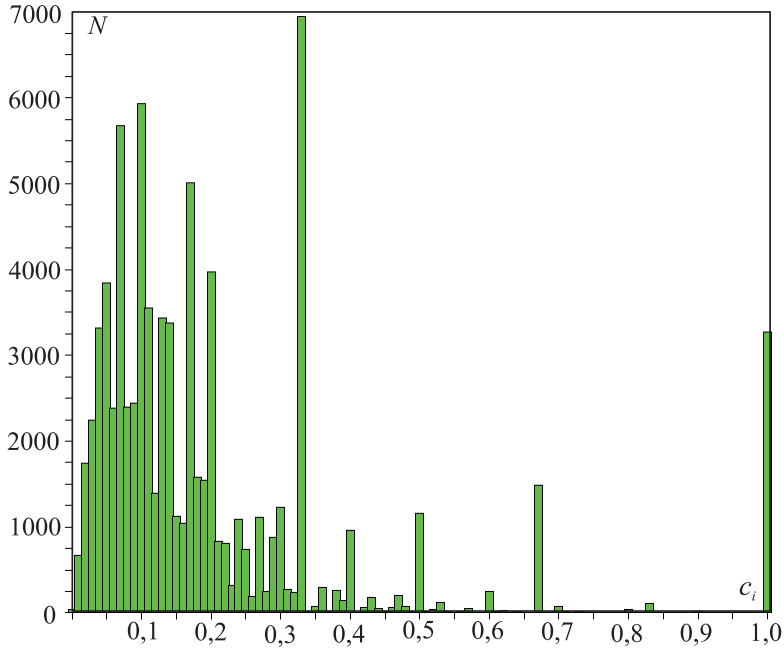


Рис. 2.1. Гистограмма локальных кластерных коэффициентов C_i графа A^u

Таблица 2.1

Кластерные коэффициенты графов A^u и A^u_{16}

A^u				A^u_{16}			
C_{avg1}	C_{avg2}	C_{rand}	$C(A^u)$	C_{avg1}	C_{avg2}	C_{rand}	$C(A^u_{16})$
0,144807	0,128781	0,000059	0,067497	0,319728	0,279762	0,242187	0,247706

одного соседа, примерно 11 % имеют двух соседей, из них у 18 % вершин соседи связаны ребром. Они имеют максимальное значение локального кластерного коэффициента, равное единице, и составляют 91 % всех вершин, имеющих максимальное значение. Для проверки зависимости C_i от степени вершины i вычислены коэффициенты корреляции Пирсона и Спирмена: $r = -0{,}042049$, $\rho = 0{,}436959$. Отрицательное значение коэффициента r указывает на обратную зависимость, т. е. большие значения C_i соответствуют небольшим значениям $deg(i)$.

Вычисление коэффициента C_{avg} выполнено в двух вариантах: а) без учета вершин, имеющих менее двух соседей; б) с учетом, при этом локальный коэффициент таких вершин считается равным нулю. Среднее значение коэффициента в первом случае обозначим C_{avg1}, во втором — C_{avg2}. Для сравнения приведем среднее значение кластерного коэффициента случайного графа $C_{rand} = deg_{avg}/n$ [10]. Вычислен также кластерный коэффициент C графа согласно (2.3). Результаты вычисления коэффициентов для компоненты A^u приведены в табл. 2.1.

Пример 2.1. Результаты вычисления значений кластерных коэффициентов для графа A^u_{16} также приведены в табл. 2.1. Отметим, что максимальное значение локального кластерного коэффициента имеет одна вершина: $C_2 = 1$. При этом ее степень $deg(2) = 2$, в то время как среднее значение степени $deg_{avg} = 3{,}9$ (см. табл. 1.1). Следующей по рангу идет вершина 16, $C_{16} = 0{,}6$. Ее степень $deg(16) = 3$. Видно отсутствие зависи-

мости между значением локального кластерного коэффициента и степенью вершины ($r = -0{,}074860, \rho = 0{,}138235$).

3. Кластерный анализ СЦС. По определению, приведенному в книге [16], процессом кластеризации называют разложение„сущностей“ в естественные группы. Столь общее определение допускает наличие неединственного способа декомпозиции исходных данных, исполняемой рекурсивно.

Полученные в результате кластеризации сетевых акторов группы рассматриваются как функциональные блоки сети и несут информацию о процессе построения сети и динамике ее роста. Решение задачи кластеризации состоит из двух этапов: а) выполнения алгоритма разбиения; б) определения качества разбиения. Результат кластеризации представляется в виде разложения множества исходных данных на подмножества. Мерой качества принято считать отношение„плотности“ элементов внутри образованных подмножеств к„разреженности“ между ними.

Итак, задача выявления сообществ акторов заключается в разбиении множества V графа $G = (V, E)$ на непересекающиеся непустые подмножества (кластеры) Cl_1,$Cl_2, \ldots, Cl_k$, так что акторы, попавшие в один кластер, расположены„ближе“ друг к другу, чем к акторам из других кластеров.

Понятие близости базируется на подобии, которое может быть основано на топологии графа, в этом случае его называют структурным. Обозначим $E(Cl_i)$ множество ребер, ориентированных из Cl_i в Cl_i [16]. Тогда $E(C) := \cup_i E(Cl_i)(i = 1, \ldots k)$ является множеством внутрикластерных ребер, а $\urcorner E(Cl) := E \backslash E(Cl)$ множеством внекластерных ребер. Кластеризация называется *тривиальной*, если $k{=}1$; *одиночной*, если $k{=}n$; *разрезом*, если $k{=}2$.

Отметим, что проблема кластеризации лучше изучена для неориентированных графов, в первую очередь это касается реализационных алгоритмов [5]. Поскольку СЦС представляет собой орграф, задача кластеризации таких объектов в интересах библиометрического анализа значительно усложняется. Существуют несколько способов применения имеющихся методик к орграфам: игнорировать ориентацию ребер; преобразовать граф в неориентированный, сохраняя информацию об ориентации, например, с помощью введения веса ребер или преобразования в двудольный граф. В работе [17] задача кластеризации формулируется как проблема минимизации суммарного веса разрезов между кластерами с целью получения кластеров сбалансированного размера. Подход позволяет применять алгоритмы кластеризации неориентированных графов для орграфов.

Простой способ игнорирования ориентации ребер мало подходит для графов цитирования, так как одновременно игнорируются семантика и приемы установления подобия согласно отношениям коцитирования или библиографического сочетания [18–20]. Примером подхода в библиометрии, заключающегося в преобразовании орграфа цитирования с матрицей смежности M в неориентированный граф, может служить построение графа коцитирования с матрицей смежности $M^{coc} = MM^\top$ или библиографического сочетания с матрицей смежности $M^{bbc} = M^\top M$. В работе [21] предлагается учесть одновременно оба отношения, рассматривая в качестве матрицы смежности графа симметричную матрицу $M^{coc+bbc} = MM^\top + M^\top M$. При таком подходе результирующий граф несет больше информации об исходном графе. Применима также технология преобразования орграфа в двудольный граф на основе понятий„центр влияния“ и„центр внимания“ [22], с последующей кластеризацией каждой доли [23, 24].

3.1. *Параметр модульность.* Совершенствование механизмов кластеризации требует развития алгоритмов оценки качества их выполнения. Общепринятым приемом является вычисление параметра модульность (*molularity*), определенного в работе [25]. В его основе лежит гипотеза о том, что структура графа, содержащего тематические сообщества, 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 случайно, независимо от типа вершин. Для неориентированного графа параметр модульность вычисляется по формуле

$$Q^u = \frac{1}{2m}\sum_{i,j}\left(m_{i,j} - \frac{k_i k_j}{2m}\right)\delta(i,j), \tag{3.1}$$

где k_i — степень вершины i; $\delta(i,j) = 1$, если i и j принадлежат одному кластеру, и $\delta(i,j) = 0$ в противном случае. Параметр модульность не подходит для сравнения результатов кластеризации графов, существенно отличающихся по размерам, он используется при принятии решений в рамках работы алгоритма кластеризации. Оптимизация модульности также используется и как метод выявления сообществ [26]. Определение модульности было расширено для орграфа в работе [27]:

$$Q^d = \frac{1}{m}\sum_{i,j}\left(m_{ij} - \frac{k_i^{out} k_j^{in}}{m}\right)\delta(i,j), \tag{3.2}$$

где k_i^{out} — исходящая степень вершины i; k_j^{in} — входящая степень вершины j.

Параметр Q фокусируется на внутрикластерных и межкластерных связях. Примером меры качества кластера, учитывающей не только количество связей между двумя группами, но и связность каждой группы со всеми вершинами графа, является мера„проводимость кластера“, определенная в работе [28].

3.2. *Мера согласованности алгоритмов кластеризации.* Для оценки результатов работы алгоритмов кластеризации используются меры эффективности, выявляющие, насколько полученные в результате работы алгоритма кластеры похожи на истинные (конечно, в предположении, что таковые известны). Такие меры, например, могут быть основаны на точности, определяющей процент акторов, правильно приписанных к кластеру, по отношению ко всем акторам или полноте, определяющей процент правильно приписанных акторов к кластеру по отношению к мощности кластера [29]. Для сравнения результатов кластеризации одного и того же множества данных различными алгоритмами также разработан ряд мер. Например, в работе [30] представлена мера *VI* (*variation of information*), определяющая количество потерянной и приобретенной информации при переходе от одного способа кластеризации к другому.

В работе [31] представлена мера *NMI* (*normalized mutual information*), определяющая степень согласованности двух делений на кластеры. Рассмотрим ее подробнее. Пусть имеются n объектов и два способа разделения на кластеры: A с кластерами C_1^A, C_2^A,...C_k^A и B с кластерами C_1^B, C_2^B,...,C_l^B. Мера *NMI* основана на матрице N^{AB} размерности $k \times l$, строки которой соответствуют номерам кластеров разделения A, столбцы — номерам кластеров разделения B, элемент $N_{i,j}^{AB} = |C_i^A \cap C_j^B|$ равен количеству объектов, общих для кластеров C_i^A и C_j^B.

$$NMI(AB) = \frac{-2\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{l} N_{i,j}^{A,B}\log\left(\dfrac{N_{i,j}^{AB}\times n}{N_i^A N_j^B}\right)}{\sum\limits_{i=1}^{k} N_i^A \log\left(\dfrac{N_i^A}{n}\right) + \sum\limits_{j=1}^{l} N_j^B \log\left(\dfrac{N_j^B}{n}\right)}, \tag{3.3}$$

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

4. Вычислительный эксперимент.

4.1. *Выявление сообществ на основе мер центральности.* Эта методика предложена в работе [32]. Алгоритм *CEB* (*community edge betweenness*, пакет *igraph*) основан на вычислении индекса центральности по посредничеству применительно к ребрам графа. Индекс учитывает долю кратчайших путей между парой вершин, проходящих через данное ребро:

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

где σ_{ij} — количество кратчайших путей от вершины i до вершины j графа, а $\sigma_{ij}(e)$ — количество кратчайших путей от i до j, проходящих через ребро e. Предполагается, что ребра между сообществами имеют большое значение индекса. Алгоритм *CEB* находит ребро с наивысшим значением индекса и удаляет его. Значение индекса для оставшихся ребер изменится. Снова вычисляется индекс и удаляется ребро с наивысшим значением. Таким образом, строится иерархический разделяющий алгоритм, который можно представить в виде дендрограммы. Для оценки качества деления на каждом уровне вычисляется параметр модульность Q (3.1, 3.2). Лучшим считается уровень, соответствующий наибольшему значению Q. Алгоритм пригоден как для неориентированных, так и для орграфов, сложность вычисления для невзвешенных графов оценивается как $\mathcal{O}(|V||E|^2)$.

Пример 4.1. Принимая во внимание высокую сложность алгоритма *CEB*, продемонстрируем два результата его работы на компоненте A_{16}. Обозначим граф, в котором ориентация ребер игнорируется, A_{16}^u, а орграф — A_{16}^d. В результате работы алгоритма для A_{16}^u получились 5 кластеров. Модульность $Q^u = 0{,}22$. Результаты работы *CEB* для A_{16}^u сравнивались с работой алгоритма *cluster_walktrap* [33], основанного на понятии свободного блуждания по графу в предположении, что сообщества представляют собой своего рода ловушки, в которых блуждание задерживается. Одним из параметров этого алгоритма является количество шагов, соответствующее глубине блуждания. На каждом шаге вычисляется модульность. В данном случае алгоритм показал одинаковое значение $Q = 0{,}26$ на шагах 2–4. Состав кластеров шага 2 практически совпадает с результатами работы алгоритма *CEB*, что соответствует $NMI = 0{,}84$.

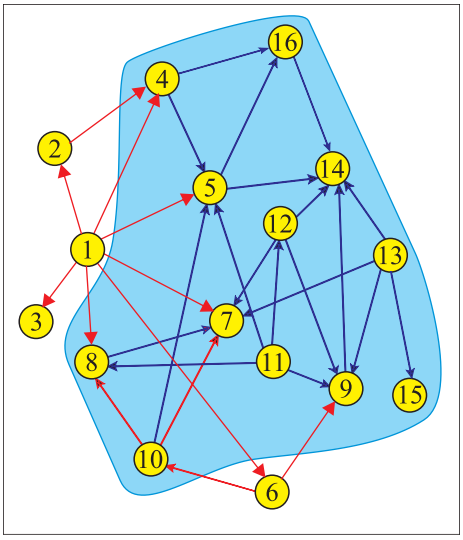


Рис. 4.1. Кластеризация графа A^d_{16} алгоритмом *CEB*

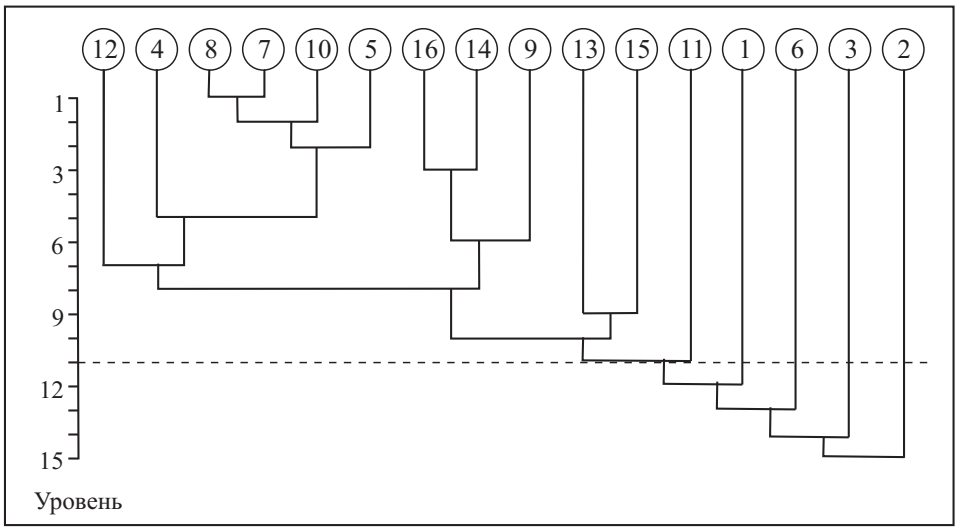


Рис. 4.2. Дендрограмма процесса кластеризации графа A^d_{16} алгоритмом *CEB*

Результаты кластеризации графа A^d_{16} алгоритмом *CEB* представлены на рис. 4.1, 4.2. Рис. 4.1 отражает разделение на кластеры. Получен один большой кластер C_1 и четыре одиночных кластера. Заметим, что C_1 не имеет связей, идущих вовне, т. е. содержит статьи, цитирующие только друг друга. Несмотря на то что получились кластеры таких разных размеров (замечание по поводу максимизации модульности, приведенное в работе [34]), в данном случае это выглядит естественно, так как отделились кластеры без внутренних связей и имеющие мало внешних связей. Исключение составляет кластер C_2, имеющий много внешних связей. На дендрограмме (рис. 4.2) представлен процесс кластеризации, пунктиром отмечен уровень 11, соответствующий максимальному значению Q^d =0,04. Согласованность результатов кластеризации графов A^u_{16} и A^d_{16} алгоритмом *CEB* соответствует NMI=0,47.

4.2. *Двухэтапная кластеризация СЦС.* К орграфам, представляющим компоненты СЦС, применен следующий метод выявления сообществ: сначала орграф преобразуется во взвешенный неориентированный, а затем используется алгоритм кластеризации, предназначенный для неориентированного графа. Рассмотрим преобразование орграфа во

взвешенные графы коцитирования A^{coc}, библиографического сочетания A^{bbc} и граф, учитывающий одновременно коцитирование и библиографическое сочетание $A^{coc+bbc}$, будем называть его суммарным. Кластеризации неориентированных графов A^{coc}, A^{bbc} и $A^{coc+bbc}$ была выполнена с помощью алгоритмов *CLP* и *CMLO* с линейной вычислительной сложностью, что позволяет применять их для больших сетевых объектов.

В работе [35] предложен алгоритм кластеризации, основанный на распространении меток вершин. На стадии инициации алгоритм присваивает всем вершинам уникальные метки. На каждой итерации узел получает метку, которую имеет большинство смежных вершин. Если имеется несколько таких меток, то выбирается произвольная. В конце работы алгоритма вершины, имеющие одинаковые метки, считаются принадлежащими одному кластеру. В идеале процесс длится до тех пор, пока находятся вершины, способные поменять метку. Реально алгоритм продолжает работу, пока каждая вершина не будет иметь такую же метку, как большинство соседей. Критерий остановки является условием, а не мерой, которую нужно максимизировать или минимизировать. В пакете *igraph* алгоритм *community_label_propagation* (*CLP*) был расширен для применения к взвешенным графам. Его сложность составляет $\mathcal{O}(|V| + |E|)$.

Алгоритм кластеризации взвешенного неориентированного графа, предложенный в работе [36], основан на оптимизации модульности за счет локальных изменений в кластерах. Он состоит из итерации двух фаз. На начальном этапе все вершины рассматриваются как кластеры. Во время первой фазы каждая вершина i перемещается в кластер смежного соседа и вычисляется модульность (3.1). Если не найдется сосед j, такой что при перемещении i в кластер, которому принадлежит вершина j, модульность разделения на кластеры будет положительной, то кластер, к которому принадлежит вершина i, не меняется. Если соседей, при перемещении к которым имеет место положительная модульность, несколько, то i перемещается в группу, обеспечивающую наибольшую модульность. Процесс последовательно повторяется для всех вершин, пока не будет достигнута наибольшая модульность. Первая фаза алгоритма закончена. Вторая фаза состоит в построении нового графа, вершинами которого являются группы, полученные во время первой фазы. Связь между двумя группами получает вес, равный сумме весов взаимных связей между вершинами групп, а сумма весов внутренних связей в группе рассматривается в качестве веса петли, замыкающейся на группу. Процесс повторяется до тех пор, пока модульность увеличивается. Алгоритм *community_multi-level_optimization* (*CMLO*) реализован в пакете *igraph*, его сложность линейна для разреженных графов.

4.3. *Результат кластеризации компоненты A.* Компонента имеет 32238 (24,5 %) изолированных вершин, не связанных отношением коцитирования, 26273 (20 %) вершин, не связанных отношением библиографического сочетания, и 2045 (1,6 %) вершин, не связанных ни одним из этих отношений. Эти вершины при кластеризации соответствующих графов алгоритмами *CLP* и *CMLO* попадают в одиночные кластеры и не учитываются при представлении результатов. Преобразуем орграф компоненты A в неориентированные графы A^{coc}, A^{bbc}, $A^{coc+bbc}$. Параметры этих графов представлены в табл. 4.1, где N_{comp} — количество компонент размером больше единицы; n_{max} — размер максимальной компоненты; Δ^{u}_{max} — ее плотность. Кластеризация выполнена с помощью алгоритмов *CLP* и *CMLO*. Результаты представлены в табл. 4.2, где N — количество кластеров, l_{avg} — средний размер кластера. Согласованность результатов, полученных с помощью алгоритмов *CLP* и *CMLO* для A^{coc}, соответствует NMI=0,81; NMI=0,77 для A^{bbc}; NMI=0,65 для $A^{coc+bbc}$. Эти значения получены с учетом одиночных кластеров. Заметим, что если при вычисле-

Таблица 4.1

Параметры графов $A^{coc}, A^{bbc}, A^{coc+bbc}$

A^{coc}			A^{bbc}			$A^{coc+bbc}$		
N_{comp}	n_{max}	Δ^u_{max}	N_{comp}	n_{max}	Δ^u_{max}	N_{comp}	n_{max}	Δ^u_{max}
622	97868	0,000357	525	104066	0,001129	219	129131	0,000916

Таблица 4.2

Кластеризация графов $A^{coc}, A^{bbc}, A^{coc+bbc}$ алгоритмами *CLP* и *CMLO*

	A^{coc}			A^{bbc}			$A^{coc+bbc}$		
	N	l_{avg}	Q (3.1)	N	l_{avg}	Q (3.1)	N	l_{avg}	Q (3.1)
CLP	1991	49,95	0,66	2137	45,65	0,66	905	104,72	0,65
CMLO	666	128,46	0,71	560	160,11	0,71	253	452,85	0,69

Таблица 4.3

Распределение размеров кластеров графов $A^{coc}, A^{bbc}, A^{coc+bbc}$. Алгоритм *CMLO*

A^{coc}		A^{bbc}		$A^{coc+bbc}$	
N	S	N	S	N	S
1	14 022	1	15 908	1	15 527
1	9689	1	9882	1	13 888
2	$8291 \le S \le 8607$	6	$7168 \le S \le 9488$	7	$7395 \le S \le 12\,675$
7	$5015 \le S \le 7052$	6	$2334 \le S \le 6044$	5	$4254 \le S \le 6199$
4	$1100 \le S \le 4575$	3	$1028 \le S \le 1548$	2	$2031 \le S \le 2580$
4	$627 \le S \le 976$	2	$514 \le S \le 529$	1	1728
2	$159 \le S \le 243$	1	211	1	833
8	$26 \le S \le 55$	3	$44 \le S \le 72$	9	$20 \le S \le 57$
83	$4 \le S \le 18$	68	$4 \le S \le 18$	21	$4 \le S \le 16$
102	3	95	3	25	3
452	2	374	2	180	2

нии параметра *NMI* исключить одиночные кластеры, соответствующие изолированным вершинам, то согласованность алгоритмов будет ниже.

Отдельно рассмотрим результаты работы алгоритма *CLMO*. Распределение размеров кластеров представлено в табл. 4.3, где N — количество кластеров, S — размер кластера. Отметим, что для всех вариантов преобразования A^{coc}, A^{bbc}, $A^{coc+bbc}$ количество кластеров размера 2 и 3 в точности совпадает с количеством компонент связности этого размера для соответствующих графов. Анализ 14 кластеров графа A^{coc} с размерами $12 \le S \le 55$ показал, что статьи, объединенные в кластеры,относятся к определенной тематике, причем в небольших кластерах более 70 % работ опубликованы в одном и том же журнале. В кластере размером 14022 максимальное число статей, опубликованных в одном журнале, составляет 4,5 %. Тематика определена по наиболее цитируемым статьям. Характеристики выборочных кластеров представлены в табл. 4.4, где S_{cl} — число статей в кластере; *Subject* — темы, объединяющие статьи; $\%J$ — наибольший процент статей, опубликованных в одном журнале; *Cit* — число цитирований наиболее популярной статьи; $Y_b - Y_e$ — год публикации самой ранней и самой поздней статей.

Пример 4.2. Рассмотрим компоненту A_{16}. Граф коцитирования A^{coc}_{16} состоит из 13 вершин, связанных 35 ребрами, и трех изолированных вершин, $\Delta(A^{coc}_{16})$ =0,29. Граф библиографического сочетания A^{bbc}_{16} состоит из 13 вершин, связанных 31 ребром, и четырех

изолированных вершин, $\Delta(A_{16}^{bbc})$ =0,26. Суммарный граф $A_{16}^{coc+bbc}$ состоит из 16 вершин, связанных 60 ребрами, $\Delta(A_{16}^{coc+bbc})$ =0,5. Кластеризация этих графов выполнена с помощью алгоритмов *CEB* и *CLMO*. Результаты работы алгоритма *CEB* для графов A_{16}^{coc} и A_{16}^{bbc} представлены на рис. 4.3, 4.4. В табл. 4.5 приведены результаты кластеризации. Согласованность результатов, полученных с помощью алгоритмов *CEB* и *CLMO*: *NMI*=0,63 для графа A_{16}^{coc}; *NMI*=0,79 для A^{bbc}; *NMI*=0,74 для $A_{16}^{coc+bbc}$. Оценка согласованности достаточно высока.

Таблица 4.4

Характеристики выборочных кластеров графов A^{coc}, A^{bbc}

A^{coc}				
S_{cl}	*Subject*	*%J*	*Cit*	$Y_b - Y_e$
1	2	3	4	5
12	Проблемы здравоохранения (healthcare problems)	83,3 %, Social Science & Medicine (Elsevier)	2	2000–2012
13	Меры продуктивности научного труда (measures of researcher's scientific output)	84,6 %, J. Informetrics (Elsevier)	8	2007–2012
14	Проблемы персонала исправительных заведений (correctional staff problems)	85,7 %, J. Criminal Justice (Elsevier)	6	1996–2008
14	Методы оценки исследований (judging research quality)	100 %, Omega (Elsevier)	10	1995–2002
14	Теория распространения слухов (theory of rumor spreading)	85,7 %, Physica A: Statistical Mechanics and its Application (Elsevier)	12	2005–2013
18	Механизмы предсказания стихийных бедствий (disaster management mechanisms)	100 %, Natural Hazards (Springer)	5	2010–2013
26	Исследование экономики Греции (analysis of Greece economics)	88,4 %, Economic Bulletin (Bank of Greece)	13	1998–2012
27	Динамика рыночных цен (dynamics of market prices)	33,3 %, J. Agricultural and Resource Economics (Wiley)	9	1982–2008
28	Теория дискретных игр (theory of discontinuous games)	21,4 %, Economic Theory (Springer); 21,4 %, J. Math. Economics (Elsevier)	8	1995–2011

Продолжение таблицы 4.4

1	2	3	4	5
31	Анализ статистических распределений (analysis of statistical distributions)	77,4 %, Annals Institute Statist. Mathem. (Springer)	16	1989–2006
32	Анализ комплексных сетей (complex networks analysis)	100 %, Physica A: Stat. Mechanics and its Applicaion (Elsevier)	11	2007–2013

36	Контрольные карты для мониторинга экон. процессов (control charts for econ. processes monitoring)	55,5 %, Intern. J. Production Economics (Elsevier)	5	1995–2014
37	Экономическое развитие в стране и за рубежом (на опыте Германии) (die wirtschaftliche entwicklung in Imland und Ausland)	32,4, RWI Konjunkturberichte (EconStor); 27 % RWI Konjunkturbericht (Rheinisch Inst.)	11	2007–2014
55	Иисследование рынка произведений искусства (art market analysis)	40 %, J. Cultural Economics (Kluwer)	19	1991–2013
627	Методы прогнозирования в экономике (automatic forecasting methods in economics)	40,4 %, J. Forecasting (Elsevier)	114	1973–2015
8607	Кредитно-денежная политика и модели (monetary policy and models)	7,8 %, J. Monetary Economics (Elsevier)	328	1994–2015
9689	Экономические модели (economic models)	7,7 % J. Banking & Finance (Elsevier)	709	1969–2015
14022	Экономические модели и эмпирические исследования (economic models and empirical investigations)	4,5 % J. Public Economics (Elsevier)	814	1960–2015
		A^{bbc}		
10	Случайное блуждание с непрерывным временем (continuous-time random walks)	50 % Statistics & Probability Letters (Elsevier) 50 % Stochastic Processes and their Applications (Elsevier)	4	2006–2013
10	Меры продуктивности научного труда (scientific impact indices)	90 %, J. Informetrics (Elsevier)	5	2008–2014
11	Анализ сейсмичности (analysis of seismicity)	100 % Physica A: Stat. Mechanics and its Applicaion (Elsevier)	5	2008–2013
11	Механизмы предсказания стихийных бедствий (natural disaster management mechanisms)	100 % Natural Hazards (Springer)	5	2011–2013
12	Анализ критических точек (change points analysis)	45,4 % Computational Statistics & Data Analysis; 36,3 % J. Multivariate Analysis (Elsevier)	2	2002–2013

Окончание таблицы 4.4

1	2	3	4	5
12	Исследование экономики Греции (Greece economics analysis)	100 %, Economic Bulletin (Bank of Greece)	9	2009–2012
14	Выполнимость центральной предельной теоремы (almost sureness of max-limit theorem)	64,3 % Statistics & Probability Letters (Elsevier)	9	1998–2013

18	Максимальные страховые выплаты и анализ точечных процессов (near-maximum insurance claims and point processes)	50 % Atatistics & Probability Letters (Elsevier)	8	1998–2012
44	Статистические распределения и их применение (statistical distributions and applications)	63,3 % Annals Inst. Math. (Springer)	5	1994–2012
50	Моделирование процессов горения каменного угля (coal combustion processes modeling)	68 % Applied Energy (Springer)	5	2011–2015
72	Моделирование работы топливных элементов (modeling of fuel cells performance)	65,3 % Applied Energy (Springer)	15	2009–2015

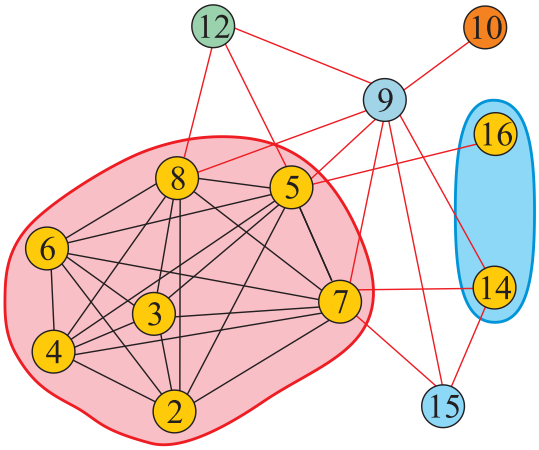


Рис. 4.3. Кластеризация графа A_{16}^{coc} алгоритмом *CEB*

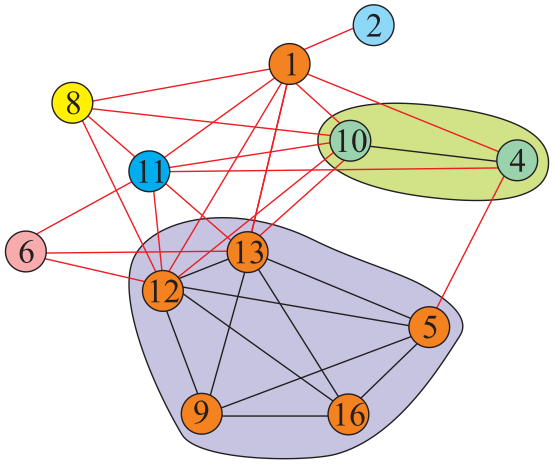


Рис. 4.4. Кластеризация графа A_{16}^{cbb} алгоритмом *CEB*

Таблица 4.5

Параметры кластеризации графов $A_{16}^{coc}, A_{16}^{bbc}, A_{16}^{coc+bbc}$ алгоритмами *CEB* и *CMLO*

Алгоритм ***CEB***						Алгоритм ***CLMO***								
A_{16}^{coc}		A_{16}^{bbc}		$A_{16}^{coc+bbc}$		A_{16}^{coc}			A_{16}^{bbc}			$A_{16}^{coc+bbc}$		
N	Q	N	Q	N	Q	N	Q	NMI	N	Q	NMI	N	Q	NMI
6	0,09	7	0,06	9	0,06	6	0,19	0,63	6	0,23	0,79	3	0,23	0,74

Заключение. Главная компонента СЦС является разреженным слабо-связным орграфом, среднее расстояние между вершинами составляет 19,7; без учета ориентации ребер составляет 16,7. Среднее значение локального кластерного коэффициента вершин значительно выше, чем у случайного графа с тем же количеством вершин (см. табл. 2.1), причем значение для вершины не зависит от ее степени. Эти параметры отражают сходство СЦС с сетевыми структурами„малого мира“, характерными для многих социальных сетей.

Выявление тематических сообществ выполнено с применением двух алгоритмов кластеризации, результаты работы которых оказались достаточно близкими по составу кластеров. Исходными данными для работы алгоритмов являлись неориентированные графы коцитирования, библиографического сочетания и„суммарный“, построенные на основе исходного орграфа.

Полученные результаты кластерного анализа позволяют аргументированно судить о направлениях научной деятельности, отраженных в статьях БД *RePEc*, периодах их публикации и цитируемости.

Список литературы

1. Бредихин С. В., Ляпунов В. М., Щербакова Н. Г., Юргенсон А. Н. Параметры "центральности" узлов сети цитирования научных статей // Проблемы информатики. 2016. № 1. С. 30–57.

2. Бредихин С. В., Ляпунов В. М., Щербакова Н. Г. Параметры пар узлов сети цитирования научных статей // Проблемы информатики. 2016. № 2. С. 30–49.

3. General principles. [Electron. resource]. http://repec.org.

4. Milgram S. The small world problem // Psychol. Today. 1967. V. 2. P. 60–67.

5. Fortunato S. Community detection in graphs // Phys. Reports. 2010. V. 486. P. 75–174.

6. Watts D. J. Small worlds: The dynamics of networks between order and randomness. Princeton: Princeton University Press, 1999.

7. Wasserman S., Faust K. Social network analysis: Methods and applications. Cambridge: Cambridge University Press, 1994.

8. Broder A., Kumar R., et al. Graph structure in the web // 9th International World Wide Web conference, Amsterdam (Netherlands), 2000. V. 33. P. 309–320.

9. Faloutsos M., Faloutsos P., Faloutsos C. On power-law relationships of the internet topology // ACM conference on applications, technologies, architectures and protocols for computer communications, Cambridge (Engl.), 1999. P. 251–262. [Electron. resource]. http://www.cs.cmu.edu/∼christos/publications/sigcomm99.pdf.

10. Watts D. J., Strogatz S. H. Collective dynamics of„small-world" networks // Nature. 1998. V. 393. P. 440–442.

11. Newman M. E. J. The structure and function of complex networks // SIAM Review. 2003. V. 45. P. 167–256.

12. Ebel H., Mielsch L. I., Bornholdt S. Scale-free topology of e-mail networks //Phys. Rev. E. 2002. V. 66, 035103.

13. Albert R., Barabasi A. L. Statistical mechanics of complex networks // Reviews of Modern Physics. 2002. V. 74. P. 47–97.

14. Burt R. S. The social structure of competition. Cambridge, MA: Harvard University Press, 1992.

15. Csárdi G., Nepusz T. The igraph software package for complex network research // InterJournal Complex Systems. 2006. 1695 P. [Electron. resource]. http://igraph.org/r/doc/.

16. Network analysis. Methodological Foundations. 2005. Springer, LNCS 3418.

17. Meilă M., Pentney W. Clustering by weighted cuts in directed graphs // Proc. of the 2007 SIAM International conference on data mining, 2007. Apr. 26–28. Minneapolis (USA). P. 135–144.

18. Маршакова И. В. Система связей между документами, построенная на основе ссылок: по данным Science Citation Index // НТИ, сер. 2. 1973. № 6. С. 3–8.

19. Small H. Co-citation in the scientific literature: A new measure of the relationship between two documents // J. Amer. Soc. Inform. Sci. 1973. V. 24, iss. 4. P. 265–269.

20. Kessler M. M. Bibliographic coupling between scientific papers // Amer. Documentation. 1963. V. 14, iss. 1. P. 10–25.

21. Satuluri V., Parthasarathy S. Symmetrizations for clustering directed graphs // Proc. 14th Internat. Conference on extending database technology, Uppsala (Sweden), March 21–25, 2011. P. 343–354. [Electron. resource]. http://dblp.uni- trier.de/db/conf/edbt/edbt2011.html.

22. Kleinberg J. M. Authoritative sources in a hyperlinked environment // J. of the ACM. 1999. V. 46, iss. 5. P. 604–632.

23. Zhou D., Schulkopf B., Hofmann T. Semi-supervised learning on directed graphs // Advances in Neural Information Processing Systems Conference, Dec. 5–8, 2005. Vancouver (Canada). P. 1633–1640.

24. Guimera R., Pardo M. S., Amaral L. A. N. Module identification in bipartite and directed networks // Phys. Rev. E 76 (3) 036102+. 2007.

25. Newman M.E.J., Girvan M. Finding and evaluating community structure in networks // Phys. Rev. 2004. E 69 (2) 026113.

26. Newman M. E. J. Fast algorithm for detecting community structure in networks // Phys. Rev. 2003. E 69 066133.

27. 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.

28. Leskovec J., Lang K. J., Dasgupta A., Mahoney M. W. Statistical properties of community structure in large social and information networks // Proc. of the 17th International Conference on World Wide Web, Beijing (China), April 21-25, 2008. P. 695–704.

29. Yang Y., Leskovec J. Overlapping community detection at scale: A nonnegative matrix factorization approach // Proc. of the 6th ACM International conference on web search and data mining, Rome (Italy). Feb. 6–8, 2013. P. 587–596.

30. Meila M. Comparing clusterings by the variation of information // Proc. of 16th annual conference on learning theory and 7th Kernel workshop, Washington (USA), Aug. 24–27, 2003. P. 173–187.

31. Fred A. L. N., Jain A. K. Robust data clustering // Proc. IEEE Computer Society conference on computer vision and pattern recognition, Minneapolis (USA), June 16–22, 2003. P. 128–136.

32. Girvan M., Newman M. E. J. Community structure in social and biological networks // Proc. Nat. Acad. Sci. USA. 2002. V. 99. P. 7821–7826.

33. Pons P., Latapy M. Computing communities in large networks using random walks // J. Graph Algorithms and Applications. 2006. V. 10, N 2. P. 191–218.

34. Chen J., Yuan B. Detecting functional modules in the yeast protein-protein interaction network // Bioinformatics. 2006. V. 22, iss. 18. P. 2283–2290.

35. Raghavan U. N., Albert R., Kumara S. Near linear time algorithm to detect community structures in large-scale networks // Phys. Rev. E 76, 036106. 2007.

36. Blondel VD., Guillaume J.L., Lambiotte R., Lefebvre E. Fast unfolding of community hierarchies in large networks // J. Stat. Mech. 2008. P10008.

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

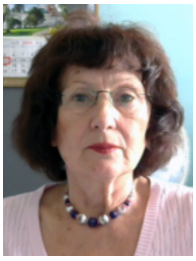
Сергей Бредихин окончил механико-математический факультет Новосибирского государственного университета в 1968 году. С 1968 года – сотрудник Института автоматики и электрометрии СО РАН. Кандидат технических наук с 1983 года. С 1988 года – заведующий Лабораторией прикладных систем Института вычислительной математики и математической геофизики СО РАН. Являлся техническим директором проекта “Сеть Интернет Новосибирского Научного Центра”. Лауреат государственной премии по науке и технике 2012 года. 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 electronnoj nauchnoy 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 года стал сотрудником Вычислительного Центра СО АН СССР, а с 1990 года – сотрудником Института систем информатики СО АН СССР. С 2004 года – 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.

Наталия Щербакова окончила Новосибирский государственный университет по специальности "Математическая лингвистика"в 1967 году. С 1967 г. работала в Институте математики СО РАН, затем в Институте автоматики и электрометрии СО РАН в области создания программного обеспечения систем передачи данных. С 2000 года – сотрудник Института вычислительной математики и математической геофизики СО РАН, где с 2002 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

Дата поступления — 09.07.2016