

# DYNAMICS OF THE CITATION NETWORK OF SCIENTIFIC ARTICLES

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

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

---

---

DOI: 10.24411/2073-0667-2020-10001

Many complex networks are scale-free, namely their degree distribution follows a power law for large  $k$ . The graphs corresponding to the citation networks ( $CN$ ) of scientific articles are included in this set. The vertices of citation network correspond to scientific articles, and directed edges to citations. Almost each new article contains some number of references (citations) to previously published ones. The number of references to a cited vertex is its in-degree. The appearance of new connects between old vertices is impossible.

The question is how growing networks self-organize into a scale-free structure. H. Simon (1955) assumes that the principle of “having much gets more” has effect. The study of this mechanism as applied to  $CN$  was performed in Price (1976). D. Price called the strategy, in which success breeds success, a *cumulative advantage*. For  $CN$ , the citation strategy is formulated as follows: the speed with which articles receive new citations is proportional to the citations already received.

Thanks to a series of works, the beginning of which was laid by the work Barabási, Albert (1999), the mechanism of cumulative advantage was called *the preferred attachment*, hereafter PA-mechanism. In Barabási – Albert model the probability  $\Pi$  that a new node connects to a node  $i$  depends on the degree  $k_i$  of  $i$  as

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}.$$

A generalization of the PA-mechanism was presented in Krapivisky, et al (2000) as

$$\Pi(k_i) = \frac{k_i^\alpha}{\sum_j k_j^\alpha} = C(t) k_i^\alpha,$$

$C(t)$  is a normalization constant. S. Dorogovtsev and J. Mendes (2000) proposed that in some real networks the probability  $\Pi$  also depends on an age of node  $i$ , decaying as  $(t - t_i)^{-\nu}$ , where  $t_i$  is a timestep a node  $i$  was added to the network,  $t$  – a timestep a new node is added,  $\nu$  is a tunable parameter. So,

$$\Pi(k_i, t_i) \propto k_i f(t_i),$$

where  $f(t_i)$  is an aging function.

In real networks  $\Pi(0) \neq 0$ , which means there is a nonzero probability that a new node attaches to a node  $i$ , such that  $k_i = 0$ . So Dorogovtsev, et al. (2000) presented the model of a directed network such that

$$\Pi(k_i^{in}) = \frac{k_i^{in} + k_0}{\sum_j (k_j^{in} + \alpha)},$$

where  $k_i^{in}$  is the number of incoming links,  $k_0$  is the initial attractiveness of the node  $i$ .

So the likelihood that a newly added node will join node  $i$  after a time  $t$  has elapsed after the addition of node  $i$  to the directed network may be proportional to the following characteristics:

$$\Pi(k_i^{in}, t) \propto (k_i^{in} + k_0) f_i(t).$$

The aim of this work is to study the dynamical properties of the citation network of scientific articles based on data provided by the bibliographic database RePEc. We find that the distribution of incoming links follows the power law with parameters  $\gamma = 2,89$ ,  $x_{\min} = 207$ . So we want to prove the preferential attachment hypothesis that in this case states that the rate  $\Pi(k^{in})$  with which a node with  $k$  incoming links acquires new links is a monotonically increasing function of  $k^{in}$ , hereinafter we use the denotation  $k$  instead of  $k^{in}$ . To investigate the attachment mechanisms separately from aging we use the method presented in Jeong, et al. (2003). To avoid the influence of  $C(t)$  and an aging effect we study the attachment kernel  $A_k$  within a relatively short window time  $\Delta T$ . The functional form of  $A_k$  can be determined by measuring how many citations an article with  $k$  citation collected during some previous period  $T_0$  receives within period  $\Delta T$ . We plotted the function of attachment rate using different previous periods and found the linear fit. The results of approximation with a linear functions show that  $A_k$  has a form  $k + k_0$ , where  $k_0$  can be considered as the initial attractiveness. It should be noted that we have different estimations of  $A_k$  depending on  $T_0$ .

When investigating an aging process we ignore the preferential attachment and explore age characteristics of nodes with the same degree. We study two distributions,  $S(t)$  and  $D(t)$ .  $S(t)$  is the distribution of ages of citation from citing article to cited articles.  $D(t)$  is the distribution of ages to a cited article from citing articles. The numerical calculations show that  $S(t)$  increases during 2–3 years and then decays exponentially during approximately 25 years (irrespective of the selected  $t_0$ ), then decreases slower. So it can be assumed that the aging function  $f(t) = (\exp \alpha t)$ ,  $\alpha < 0$ . The distribution  $D(t)$  shows the linear growth for 20 years and then the exponential decay.

Our experiment shows that the citation process can be described by the linear preferential attachment, but given the process of aging is consistent with the attachment kernel only for the first 20 years after publication. It should be noted that fluctuations in the distributions and the dependence on selected time windows are observed which may be due to incomplete data.

**Key words:** citation network, power law, preferential attachment, initial attractiveness, aging.

## References

1. LOTKA A. J. The frequency distribution of scientific productivity // J. of the Washington Academy of Science. 1926, V. 16. P. 675–682.
2. SIMON H. A. On a class of skew distribution functions // Biometrika. 1955. V. 42. P. 425–440.
3. PRICE D. J. DE SOLLA. Networks of Scientific Papers // Science. 1965. V. 149. P. 510–515.
4. MERTON R. K. The Matthew Effect in Science // Science. 1968. V. 159, iss. 3810. P. 56–63.
5. COLE J. R., COLE S. Social stratification in science. Chicago, IL: University of Chicago. 1973.
6. PRICE D. J. DE SOLLA. A general theory of bibliometric and other cumulative advantage processes // J. of the American Society for Information Science. 1976. V. 27(5–5). P. 292–306.
7. TSALLIS C., DE ALBUQUERQUE M. P. Are citations of scientific papers a case of nonextensivity? // Eur. Phys. J. B. 2000. V. 13, iss. 4. P. 777–780.
8. REDNER S. How popular is your paper? An empirical study of the citation distribution // Eur. Phys. J. B. 1998. V. 4, iss. 2. P. 131–134.

9. PETERSON G. J., PRESSE S., DILL K. A. Nonuniversal power law scaling in the probability distribution of scientific citations // in Proc. Natl. Acad. Sci. USA. 2010. V. 107, iss. 37. P. 16023–16027.
10. BARABASI A.-L., ALBERT R. Emergence of scaling in random networks // Science. 1999. V. 286. P. 509–512.
11. REPEC. General principles. [Electron. Resource]. <http://repec.org/>.
12. KRAPIVISKY P. L., REDNER S., LEYVRAZ F. Connectivity of growing random networks // Phys. Rev. Lett. 2000. V. 85. P. 4629–4632.
13. SCHERBAKOVA N. G. Preferential attachment models // Problemi informatiki. 2019. N 3. P. 46–61 (in Russian).
14. BARABASI A.-L., ALBERT R., JEONG H. Mean-field theory for scale-free random networks // Physica A. 1999. V. 272. P. 173–187.
15. DOROGOVTSSEV S. N., MENDES J. F. F., SAMUKHIN A. N. Structure of growing network with preferential linking // Phys. Rev. Lett. 2000. V. 85. P. 4633–4636.
16. BARABASI A. L., JEONG H., NEDA Z., RAVASZ E., SCHUBERT A., VICSEK T. Evolution of the social network of scientific collaborations // Physica A. 2002. V. 311. P. 590–614.
17. JEONG H., NEDA Z., BARABASI A. L. Measuring preferential attachment for evolving networks // EuroPhysics Letters. 2003. V. 61. P. 567–572.
18. Lehmann S., Lautrup B., Jackson A. D. Citation networks in high energy physics // Phys. Rev. E. 2003. V. 68, 026113.
19. REDNER S. Citation statistics from more than the century of physical review // arXiv:physics/0407137.
20. WANG M., YU G., YU D. Measuring the preferential attachment mechanism in citation networks // Physica A. 2008. V. 387. P. 4692–4698.
21. DOROGOVTSSEV S. N., MENDES J. F. F. Evolution of reference networks with aging // Phys. Rev. E. 2000. V. 62. P. 1842–1845.
22. ZHU H., WANG X., ZHU J.-Y. Effect of aging on network structure // Phys. Rev. E. 2003. V. 68, 056121.
23. HAJRA K.B., SEN P. Modelling aging characteristics in citation networks // Physica A. 2006. V. 368. P. 575–582.
24. CLAUSET A., SHALIZI C. R., NEWMAN M. E. J. Power-law distributions in empirical data // SIAM Review V. 51. 2009. P. 661–703.
25. IGRAPH – The network analysis package [Electron. Resource]. <https://igraph.org/>.
26. POLLMANN T. Forgetting and the ageing of scientific publications // Scientometrics. 2000. V. 47, N 1. P. 43–54.
27. LEHMANN S., JACKSON A. D., LAUTRUP B. Life, death and preferential attachment // EuroPhysics Letters. 2005. V. 69. P. 298–303.

## ДИНАМИКА РОСТА 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

DOI: 10.24411/2073-0667-2020-10001

Приведены результаты эмпирического исследования параметров процессов, обеспечивающих динамику развития сети цитирования статей: предпочтительное присоединение, старение информации и начальная привлекательность статей. Измерена скорость, с которой статьи получают новые цитирования, и показана ее линейная зависимость от числа уже имеющихся цитирований. Также измерена скорость “старения” статей, влияющая на процесс получения цитирований. Приведена оценка параметра “начальная привлекательность” узлов.

**Ключевые слова:** сеть цитирования статей, степенной закон, процесс предпочтительного присоединения, первоначальная привлекательность узла, процесс старения узла.

**Введение.** Комплексные сети охватывают обширное семейство сетей, представляющих различные области деятельности: социальные сети, WWW, электронная почта и т. п. Графы этого семейства характеризуются большим числом вершин и невысокой плотностью ребер. Наиболее изученные классы комплексных сетей – это сети “малого мира”, которым свойственны небольшое среднее расстояние между связными узлами и высокий коэффициент кластеризации, а также безмасштабные сети, с распределением степеней узлов, следующим степенному закону. Сети цитирования научных статей (далее СЦС) входят в класс безмасштабных сетей. Они обладают важным свойством роста: добавление “новых” узлов (статей) и установление “новых” ребер (цитирований) между “старыми” и “новыми” узлами приводит исключительно к росту структуры графа, поскольку “старые” узлы и ребра не удаляются.

Анализ степенных распределений в библиометрии начат в работе [1], в которой было построено распределение авторов химических рефератов за период 1907–1916 гг. и эмпирическим путем получен вывод о том, что число авторов, опубликовавших  $n$  статей, пропорционально  $1/n^2$ . Значительно позже автор работы [2] показал, что хвосты асимметричных распределений могут быть достаточно точно аппроксимированы функцией вида  $f(x) = (a/x^k)b^x$ , где  $a, b, k$  – константы, зависящие от типа данных. В работе [3] установлено, что доля статей, получивших  $k$  цитирований (для достаточно больших  $k$ ), уменьшается пропорционально  $k^{-\gamma}$ , где  $2 \leq \gamma \leq 3$ . То есть вероятность того, что узел имеет входящую степень, равную  $k$ , подчиняется степенному закону  $P(x) \sim x^{-\gamma}$ , где  $\gamma > 1$ ,  $x > x_{\min}$ .

Было замечено, что степенной закон возникает, когда выполняется принцип “имеющий много получает больше”. В социологии этот принцип называется “эффектом Матфея” (термин предложен Р. Мертоном в работе [4] и основан на библейском изречении). Влиянию

этого процесса на распределение ресурсов и социальное расслоение общества посвящена монография [5]. В библиометрии этот процесс выглядит следующим образом: авторитетный ученый, имеющий существенный кредит доверия, получает на свои статьи большее число цитирований, нежели менее авторитетный, что повышает его статус.

Исследование этого процесса применительно к СЦС выполнено в работе [6]. Автор представил процесс цитирования, при котором успех порождает успех и назвал его *кумулятивным преимуществом*. Для СЦС стратегия цитирования формулируется так: скорость, с которой статьи получают новые цитирования, пропорциональна уже полученным цитированиям. В качестве подтверждения того, что такая стратегия ведет к степенному закону распределения входящих цитирований, рассмотрена модель ориентированной сети — модель Прайса. Рост сети обеспечивается добавлением новых узлов не обязательно с постоянной частотой. Новые узлы имеют различную исходящую степень (число цитируемых статей), но средняя исходящая степень является константой в рассматриваемый временной период и обозначается  $m$ . Соответственно, средняя входящая степень узла также равна  $m$ . Доля узлов с входящей степенью  $k$  обозначается  $p_k$ , а  $p_{k,n}$  — доля таких узлов в сети, имеющей  $n$  узлов. На старте каждый узел имеет нулевую степень, поэтому предполагалось, что вероятность присоединения к узлу пропорциональна  $k + k_0$  (в модели Прайса  $k_0 = 1$ ). Показано, что  $p_k \sim k^{-(2+1/m)}$ , значение  $\gamma$  изменяется в интервале (2, 3).

В работах [7–9] приведены результаты анализа распределения степеней узлов реальных сетей цитирования, подтверждающие следование степенному закону. Сети с распределением степеней узлов, подчиняющихся степенному закону, в работе [10] получили название *безмасштабные*.

Значительная часть работы посвящена вычислительному эксперименту, направленному на измерение параметров роста библиографической БД. В результате оценена вероятность появления новых узлов и ребер, исследован процесс “старения” (потери интереса к содержанию статьи). Эксперимент выполнен на данных, извлеченных из БД RePEc [11].

**1. Процесс предпочтительного присоединения.** Благодаря серии работ, начало которой положено работой [10], процесс, позволяющий объяснить топологические свойства комплексных сетей, получил название *предпочтительное присоединение* (далее РА). В работе [10] представлена модель (далее ВА-модель), позволяющая строить безмасштабные неориентированные сети. На старте имеется  $m_0$  произвольно связанных узлов, каждый из которых имеет хотя бы единичную степень. Алгоритм развития сети включает два этапа. На первом (этап роста) в каждый момент времени к сети добавляется новый узел, имеющий  $m < m_0$  ребер для присоединения к  $m$  различным уже имеющимся узлам. На втором этапе благодаря РА добавленный узел присоединяется к узлу  $i$  с вероятностью

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}, \quad (1)$$

здесь  $k_i$  — степень узла  $i$ .

Согласно ВА-модели, после  $t$  шагов сеть будет иметь  $n = t + m_0$  узлов и  $m_0 + mt$  ребер. В предположении, что  $k_i$  — непрерывная действительная переменная, эволюцию степени  $k_i$  узла  $i$  можно представить уравнением

$$\frac{dk_i}{dt} = m\Pi(k_i). \quad (2)$$

Построенная сеть имеет распределение степеней узлов, следующее степенному закону с экспонентой  $\gamma = 3$ .

Обобщение РА представлено в работе [12]:

$$\Pi(k_i) = \frac{k_i^\alpha}{\sum_j k_j^\alpha} = C(t)k_i^\alpha, \quad (3)$$

где  $C(t)$  — нормализующая константа, зависящая от времени, а вид функции распределения степеней узлов зависит от значения  $\alpha$ . Для  $\alpha < 1$  распределение имеет вид  $P(k) \sim e^{k^{-\gamma}}$ ; для  $\alpha > 1$  возникает феномен “застывания”, когда один узел оказывается связанным почти со всеми остальными. Только в случае линейной зависимости,  $\alpha = 1$  (см. (1)), сеть становится безмасштабной. Обзор процессов, определяющих структуру комплексных сетей, представлен в работе [13].

Для анализа ВА-модели и ее расширений, представленных в работах [12, 14, 15], предложены аналитические методы, позволяющие характеризовать динамику развития сети. Эти методы также применяются при проведении измерений в реальных сетях. Измерения в безмасштабных сетях подтверждают гипотезу о том, что плотно связанные узлы наращивают связность быстрее, чем менее связанные. Результаты соответствующих статистических исследований приведены в работах [16–20].

Для извлечения функциональной формы  $\Pi(k_i)$  на основе реальных данных в работе [17] представлена следующая методика. С учетом (2), (3) вероятность  $\Pi(k_i)$  рассматривается как скорость, с которой существующий узел  $i$  степени  $k$  приобретает новые ребра в процессе роста сети. То есть следует определить число приобретенных ребер как функцию от степени узла. Поскольку константа  $C(t)$  в (3) зависит от времени присоединения к сети узла  $i$ , выбираются достаточно короткие интервалы, в течение которых происходит присоединение новых узлов. Рассматривается сеть, для которой известен порядок появления узлов и ребер, например библиометрическая сеть. Множество узлов, присутствующих на момент  $T_0$ , назовем  $T_0$ -узлами. Множество  $T_1$ -узлов — это узлы, добавленные в интервал времени  $d = [T_1, T_1 + \Delta T]$ ,  $|T_1| > |T_0|$ ,  $\Delta T \ll T_1$ . Когда к сети добавляется  $T_1$ -узел, вычисляем, какова степень  $T_0$ -узла, с которым соединяется  $T_1$ -узел. Гистограмма, отображающая число ребер, приобретенных  $T_0$ -узлами, имеющими степень  $k$ , после нормализации  $(\Delta k_i / \Delta T)$ , задает скорость  $\Pi_d(k)$  на интервале  $d$ . Если сеть развивается стационарно, т. е.  $\Pi_d(k)$  зависит не от выбора интервала, а только от  $k$ , то  $\Pi_d(k)$  соответствует функции предпочтительного присоединения. В работах [16, 17] с целью сглаживания колебания в рамках разных временных отрезков исследуется кумулятивная функция:

$$c(k) = \int_0^k \Pi(k) dk.$$

Если  $\Pi(k)$  следует (3), то  $c(k) \propto k^{\alpha+1}$ .

**2. Процесс старения узла.** Для СЦС возраст статьи является важным фактором, влияющим на эволюцию сети. Влияние проявляется в затухании интереса к ранним статьям, которые перестают получать цитирования. Для этого есть несколько причин, например, представленные в них идеи могут быть развиты в более поздних работах, которые и получают цитирования. Кроме того, число статей неуклонно растет, отнимая внимание у более ранних. “Большинство цитирований относится к свежим статьям, так как большинство статей являются свежими” [6]. Конечно, это не касается выдающихся статей, имеющих стабильно высокий уровень цитирования. Аналитические исследования показали, что вид функции старения существенно влияет на распределение степеней узлов. В

работе [21] рассмотрена модель, являющаяся расширением ВА-модели, в которой вероятность присоединения нового узла к узлу  $i$  зависит не только от степени узла  $k_i$ , но и от возраста. Если  $t_i$  — момент присоединения к сети узла  $i$ ,  $t$  — текущий момент, то

$$\Pi(k_i, t_i) \propto k_i f(t_i), \quad (4)$$

где  $f(t_i)$  — функция старения. Функция старения определена как степенная функция  $f(t_i) = \tau^{-\nu}$ ,  $\tau = (t - t_i)$  — возраст статьи,  $\nu$  — параметр затухания. Показано, что это изменение ВА-модели является критичным: если  $\nu > 1$ , распределение степеней экспоненциальное, а если  $\nu < 1$  — степенное и параметр  $\gamma$  зависит от значения  $\nu$ .

Экспоненциальный вид функции  $f(t_i)$  рассматривается в работах [22, 23]. В первой из них изучается влияние затухания вида  $e^{-\beta\tau}$  на параметры сети — кластерный коэффициент и среднее расстояние между узлами. Во второй работе рассматриваются две модели, для которых вероятность присоединения определяется как  $\Pi(k, t) \sim k^\beta t^\alpha$  и  $\Pi(k, t) \sim k^\beta \exp(\alpha t)$  и исследуется, при каких значениях параметров  $\beta$  и  $\alpha$  сеть является безмасштабной.

**3. Параметр „начальная привлекательность“.** Расширение ВА-модели для ориентированных сетей приведено в работе [15]. В каждый момент к сети добавляется новый узел  $j$ , имеющий  $m$  ориентированных ребер, который присоединится к узлу  $i$  (в сети появится дуга  $(j, i)$ ) с вероятностью

$$\Pi(k_i^{in}) = \frac{k_i^{in} + k_0}{\sum_j (k_j^{in} + a)}, \quad (5)$$

где  $k_i^{in}$  — входящая степень узла  $i$ ,  $k_0 \geq 0$  — одинаковая для всех узлов константа, называемая “начальная привлекательность”. Наличие  $k_0$  позволяет “молодым” узлам получать новые ребра. Для модели, приведенной в работе [15], исходящая степень всех узлов одинакова и равна  $m$ . Степень узла  $k = k_i^{in} + m$ , т.е., если  $k_0 = m$ , то модель совпадает с ВА-моделью. Интерес представляет распределение входящих степеней узлов. В работе [15] показано, что функция распределения входящих степеней следует степенному закону, а экспонента  $\gamma$  зависит от  $k_0$ .

Принимая во внимания (1), (4), (5), предполагаем, что вероятность присоединения к узлу  $i$  вновь добавляемого узла по прошествии времени  $t$  после появления в сети узла  $i$  пропорциональна

$$\Pi(k_i^{in}, t) \propto (k_i^{in} + k_0) f_i(t). \quad (6)$$

**4. Вычислительный эксперимент.** Узлами СЦС являются статьи, а упорядоченные пары статей, связанные отношением цитирования — ориентированными ребрами. Далее обсуждаются результаты анализа реальной сети цитирования, проведенного с целью определить, как факторы, указанные в (6), влияют на динамику развития сети. Заметим, что эти факторы не исчерпывают все предпосылки, влияющие на получение цитирований, такие как нормы цитирования, обоснованность ссылок, личные предпочтения и т. д.

Особое внимание уделяется РА, являющемуся ядром процесса аккумуляции цитирований. Необходимо выяснить, зависит ли вероятность получения новых цитирований от количества уже полученных и какова функциональная форма этой зависимости. В ходе изложения будем пользоваться следующей терминологией. Если цитирующая статья в списке литературы содержит указание на какую-либо статью, то такая

Таблица 1

Средний возраст цитирования статей

Число цитирований	Число цитируемых статей	$\langle Age \rangle$
> 500	301	22,87
> 200	1677	19,65
> 100	5311	17,41
> 50	15 412	15,47
> 20	53 265	13,48
< 10	501 287	7,52
< 5	388 583	7,23
= 1	183 145	7,18

ссылка с точки зрения цитирующей статьи называется *исходящим цитированием*, а с точки зрения цитируемой статьи — *входящим цитированием*. Пусть статья  $j$  цитирует статью  $i$ . Обозначим  $t_j$  и  $t_i$  — годы публикации статей  $j$  и  $i$ . *Возраст цитирования*  $Age$  статьи  $i$  статьей  $j$  определяется разностью  $\tau_i = t_j - t_i$ .

4.1. *Исходные данные.* Данные о цитировании журнальных статей за период с 1874 г. по 30 июня 2019 г. извлечены из библиографической БД RePEc. Всего статей 1 404 431. Отсеяны статьи, в идентификаторе которых отсутствует явное указание года, имеются ссылки на неправильный год и самоцитирования. Исследуется максимальная компонента  $N_{REP} = (V, E)$  ориентированной СЦС, содержащая  $|V| = 819\,207$  узлов и  $|E| = 5\,538\,043$  ребер. Учитываются только внутренние цитирования, т.е. цитирования между статьями, содержащимися в  $V$ . Статей, не имеющих цитирований, — 191 468 (23,37%). Статей, процитированных хотя бы один раз, — 627 739. Среднее число входящих цитирований для статей множества  $V$  составляет 6,76. Для статей, процитированных хотя бы один раз, — 8,82. Табл. 1 позволяет оценить средний возраст цитирования статей в соответствии с определенными диапазонами числа цитирований.

Средний возраст цитирования  $\langle Age \rangle = 11,02$ . Наиболее цитируемые статьи имеют возраст цитирования значительно выше среднего. Для статей, имеющих более 500 цитирований, средний возраст цитирования  $\langle Age \rangle = 22,87$ .

Для  $N_{REP}$  характерен экспоненциальный рост числа статей и, соответственно, новых исходящих цитирований. Например, в 2000 г. БД пополнилась на 15 445 статей и 57 086 исходящих цитирований (3,69 ссылки на одну статью), а в 2010 г. — на 35 406 статей и 249 324 новых цитирований (7,04 ссылки на статью). Увеличилось и число входящих цитирований, относящихся к статьям, опубликованным за эти годы: 175 578 и 242 655 соответственно.

На рис. 1 представлено распределение входящих и исходящих цитирований по годам с 1937-го по 2019-й. По оси абсцисс указаны годы с шагом 10 лет, а по оси ординат — число (степени десятки) входящих (синий цвет) и исходящих (красный цвет) цитирований. Кривые имеют одинаковую тенденцию роста. К ранним годам относится незначительное число статей и, соответственно, исходящих цитирования. Близость кривых показывает, что в основном цитируются недавно вышедшие статьи. Отметим, что множество содержит всего 433 статьи, опубликованных до 1936 г.

На рис. 2 представлено распределение входящих степеней узлов для множества  $V$  и для статей, относящихся к выбранным годам (см. правый верхний угол). По оси абсцисс показано число входящих цитирований  $k$ , а по оси ординат — доля статей, имеющих  $k$

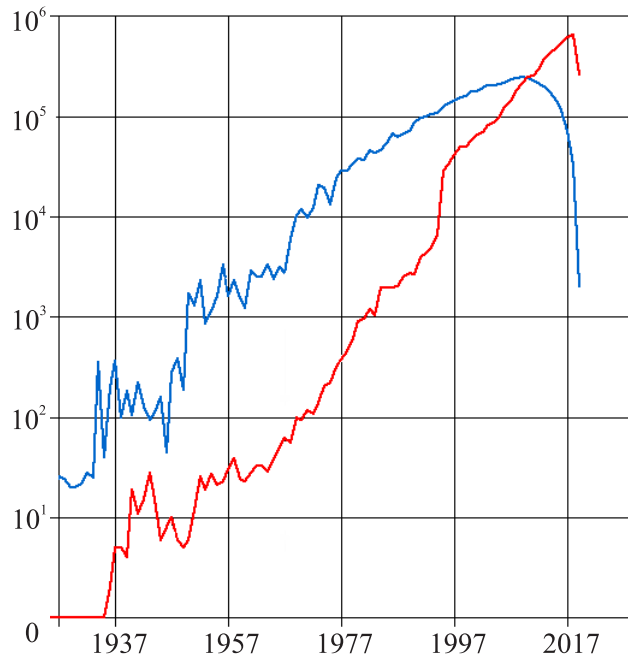


Рис. 1. Число цитирований по годам

цитирований. Обе оси имеют логарифмический масштаб. Видно, что графики подобны, за исключением больших значений  $k$ . Близость распределений указывает на то, что зависимость от роста числа статей мала.

Метод распознавания наличия степенного закона в эмпирических данных и определения значений параметров предложен в работе [24]. С использованием пакета *igraph* [25] установлено, что распределение входящих степеней узлов следует степенному закону с параметрами  $\gamma = 2,89$ ,  $x_{\min} = 207$ . Таким образом, сеть  $N_{REP}$  является безмасштабной.

4.2. *Скорость присоединения.* Исследуем (6) независимо от  $k_0$  и  $f_i(t)$ , зафиксировав короткое временное окно. Рассмотрим изменение входящей степени узлов сети  $N_{REP}$ . Скорость, с которой ранее опубликованная статья, имеющая  $k$  цитирований, будет процитирована вновь опубликованной, обозначим  $A(k)$ , это — основа процесса присоединения. Игнорируем константу  $C(t)$  и покажем, что для рассматриваемой сети  $A(k) \propto k$ .

Следуя методике, изложенной в работах [17, 19], определим  $T_1 = 2018$ ,  $\Delta T = 1$  и рассмотрим исходящие цитирования статей, изданных в период  $[T_1, T_1 + \Delta T]$ , т. е. в течение 2018 г. Число исходящих цитирований за этот период составляет 658 415. Пусть  $T_0 = T_1 - 1$ , т. е.  $T_0 = 2017$ . Для каждой статьи, опубликованной в период  $w_1 = [1874, T_0]$ , подсчитываем общее число входящих цитирований  $k$ , полученных к моменту  $T_1$ . Подсчитываем  $\Delta k$  — среднее число входящих цитирований, полученных статьями с  $k$  цитированиями в период  $[T_1, T_1 + \Delta T]$ . Рассматриваем дополнительные окна  $w_2 = [1988, T_0]$ ,  $w_3 = [1998, T_0]$ ,  $w_4 = [2008, T_0]$ , т. е. меняем начальный год периода, для которого подсчитывается  $k$ .

На рис. 3 приведены зависимости  $\Delta k$  от  $k$  для окон  $w_2 - w_4$ . Представлены значения  $k \leq 1400$ , поскольку доля статей, для которых  $k > 1400$ , составляет 0,0064% от общего числа. Видно, что зависимость графиков от рассматриваемого окна не существенна.

Для аппроксимации функциональной формы  $A(k)$  линейной функцией вида  $y = ax + b$  использовался метод наименьших квадратов. Заметим, что для малых значений  $k$  линей-

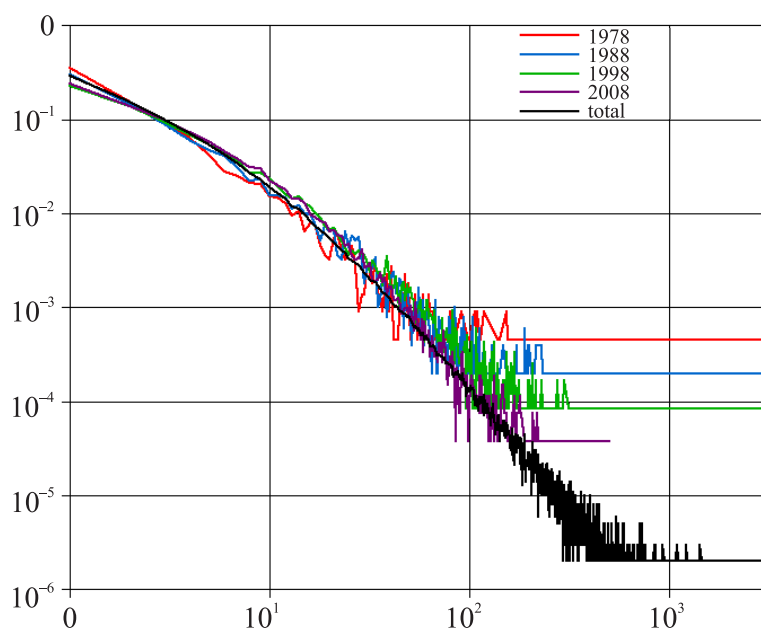


Рис. 2. Распределение входящих степеней узлов

Таблица 2

Аппроксимация  $A(k)$ 

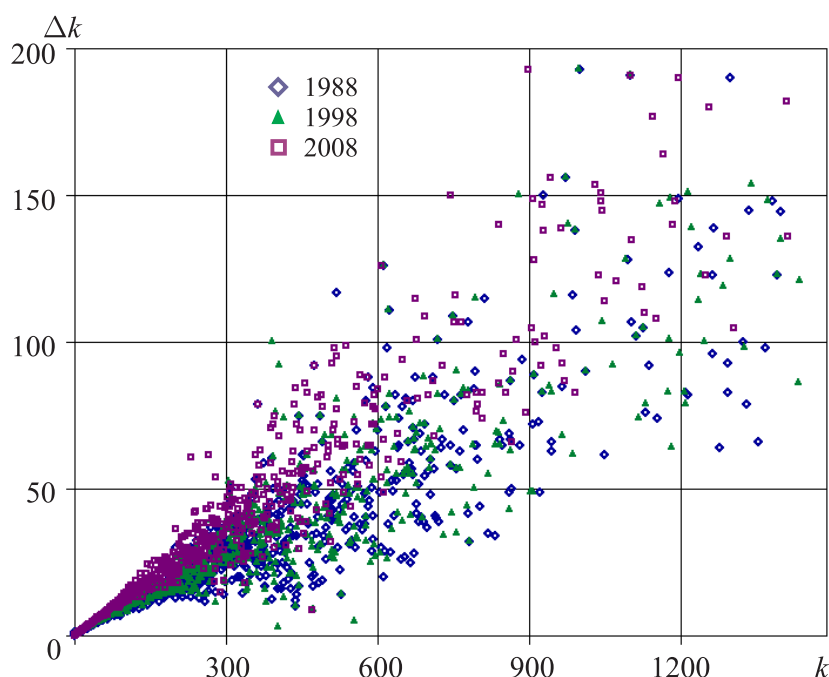
Окно	Функция 1	Функция 2
$w_1$	$y = 0,090066x - 0,404866$	$y = 0,094525x + 0,369329$
$w_2$	$y = 0,090105x - 0,405038$	$y = 0,094847x + 0,365893$
$w_3$	$y = 0,095328x - 0,959371$	$y = 0,098202x + 0,345890$
$w_4$	$y = 0,126907x - 0,682570$	$y = 0,127795x + 0,213792$

ность  $A(k)$  прослеживается визуально. В качестве примера см. рис. 4, на котором представлен график функции  $A(k)$ ,  $k < 300$ , для периода  $w_3$ . Вид функции, аппроксимирующей  $A(k)$  для рассмотренных временных окон, приведен в табл. 2, столбец 2.

Представим выражение  $ax + b$  в виде  $x + b/a$ . Тогда  $A(k) \propto k + k_0$ , где  $k_0$  можно рассматривать как начальную привлекательность. В нашем случае  $k_0$  имеет отрицательное значение для всех окон. Так,  $k_0 \sim -4,495$  для  $w_1$  и  $w_2$ . Аппроксимируем  $A(k)$  для  $k \leq 50$  (см. табл. 2, столбец 3). В этом случае  $k_0 > 0$ , причем для окон  $w_1 - w_3$  имеет близкие значения (3,9–3,5). Поскольку длина периода  $\Delta T$  мала, оценка нестабильна. Выбор нескольких временных периодов для вычисления  $k$  позволяет дать агрегированную оценку, но при этом игнорируется изменение нормализующей константы.

4.3. *Процесс старения.* Рассмотрим влияние эффекта старения на эволюцию сети  $N_{REP}$ , т.е. исследуем функцию  $f_i(t)$ . Так же как в п. 4.2, изучаем эффект, игнорируя РА. Вычисляем, через какой промежуток времени статьи с одним и тем же числом цитирований будут вновь процитированы.

В результате анализа зависимости между средним возрастом цитирования и числом получаемых цитирований выявлена положительная корреляция, которая прослеживается для числа цитирований меньше 50 (аппроксимация  $y = 0,08882x + 7,3354$ ). Для больших значений наблюдаются значительные флуктуации, причем амплитуда колебаний увели-

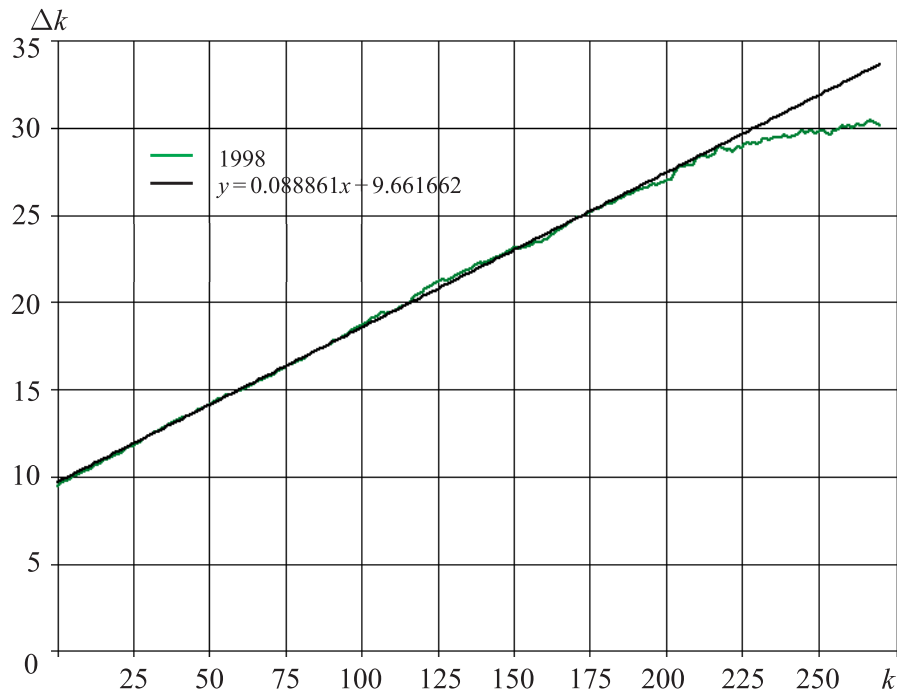
Рис. 3. Скорость получения цитирований  $A_k$ 

чивается. “Выдающиеся” (часто цитируемые) статьи имеют большой средний возраст цитирования. То есть процессы увеличения возраста статей и старения различны.

Отметим, что некоторые статьи выделяются из общей картины. Это статьи, окончательно *утратившие внимание*. Следуя работе [19], такими будем считать статьи, имеющие к 2019 г. меньше 50 входящих цитирований, у которых средний возраст цитирования меньше, чем  $1/3$  возраста самих статей, дополнительно включаем только статьи, опубликованные ранее 2000 г. Таких оказалось 3,8 % (87 047) от всех процитированных хотя бы один раз, причем большинство из них (62 258, 71,5 %) остались неактивными до конца всего периода рассмотрения ( $w_1$ ). На наличие таких статей указывается, например, в работах [19, 26, 27].

Теперь рассмотрим распределение возрастов цитируемых статей относительно цитирующих. Зафиксируем  $t_0$  — год публикации цитирующих статей. Статьи, опубликованные в году  $t_0$ , цитируют  $n$  статей, опубликованных в годах  $t_1, t_2, \dots, t_j, \dots$  (в прошлом относительно  $t_0$ ). Пусть  $n_i$  — число цитируемых статей, опубликованных в году  $t_i$ . Распределение интервалов  $\theta_i = (t_0 - t_i)$  задает  $S(t)$ .

На рис. 5 представлено распределение возрастов цитируемых статей, опубликованных в период  $[1950, t_0]$ . Для более длинного интервала данных о цитировании в БД недостаточно, поскольку таких публикаций мало, их исключение влияет незначительно. На оси абсцисс указаны интервалы времени, прошедшего с момента публикации цитируемых статей, а на оси ординат — доля цитируемых статей  $n_i/n$ , относящихся к этому интервалу (шкала логарифмическая). Независимо от выбора  $t_0$ , графики демонстрируют одни и те же темпы потери интереса к “старым” статьям, а поскольку количество “старых” статей мало, то суммарное распределение также незначительно отличается от приведенных на рисунке. Графики демонстрируют интерес к статьям, опубликованным за 2–3 года до публикации цитирующей статьи, затем идет экспоненциальное падение интереса, соответствующее  $y =$

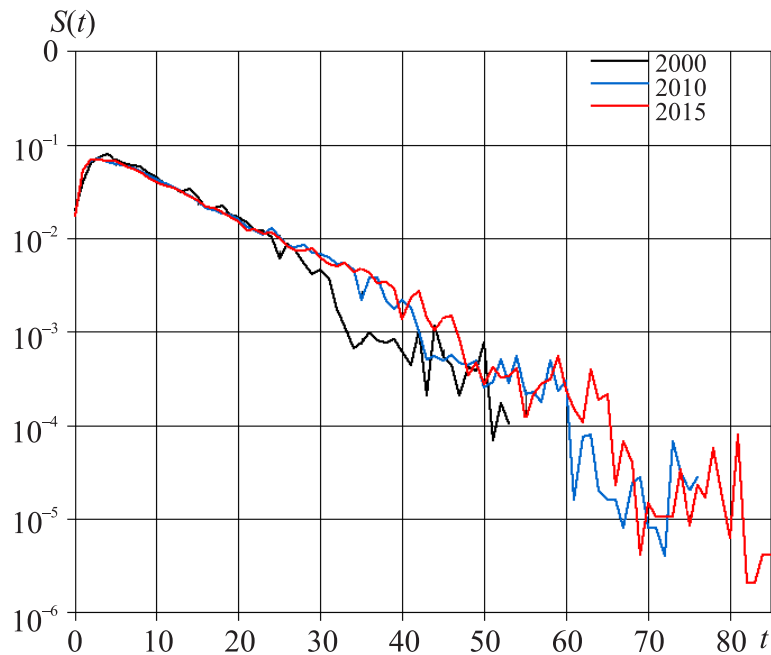
Рис. 4. Скорость получения цитирований для малых  $k$ 

$\exp(\lambda x)$ . Около 25 лет затухание идет наиболее быстрыми темпами. Так, для 2015 г. в период от 3 до 24 лет  $\lambda = -0,093$ , а в период от 25 до 70 лет  $\lambda = -0,118$ . В целом  $S(t)$  имеет экспоненциальное убывание для небольших значений  $t$ . Распределение  $S(t)$  является аналогом  $f(t)$  из (6), т.е. функция затухания имеет вид  $f(t) = \exp(\alpha t)$ ,  $\alpha < 0$ .

В работе [21] показано, что для модели сети с постоянной скоростью роста экспоненциальное затухание ведет к экспоненциальному распределению степеней узлов. Однако из п. 4.1 следует, что в сети  $N_{REP}$  распределение входящих степеней узлов подчиняется степенному закону. Подчеркнем, что эта сеть растет экспоненциально и, как замечено в работе [19], в этом случае на нее в меньшей степени влияет быстрое затухание интереса к “старым” статьям.

Рассмотрим распределение  $D(t)$  возрастов цитирующих статей относительно цитируемых, опубликованных в определенном году. Фиксируется  $t_0$  — год публикации статей, для которых рассматриваются входящие цитирования, они поступают от  $n$  цитирующих статей, опубликованных в годах  $t_1, t_2, \dots, t_j, \dots$  (в будущем относительно  $t_0$ ). Пусть  $n_i$  — число цитирующих статей, опубликованных в году  $t_i$ . Распределение интервалов  $\theta_i = (t_i - t_0)$  задает  $D(t)$ .

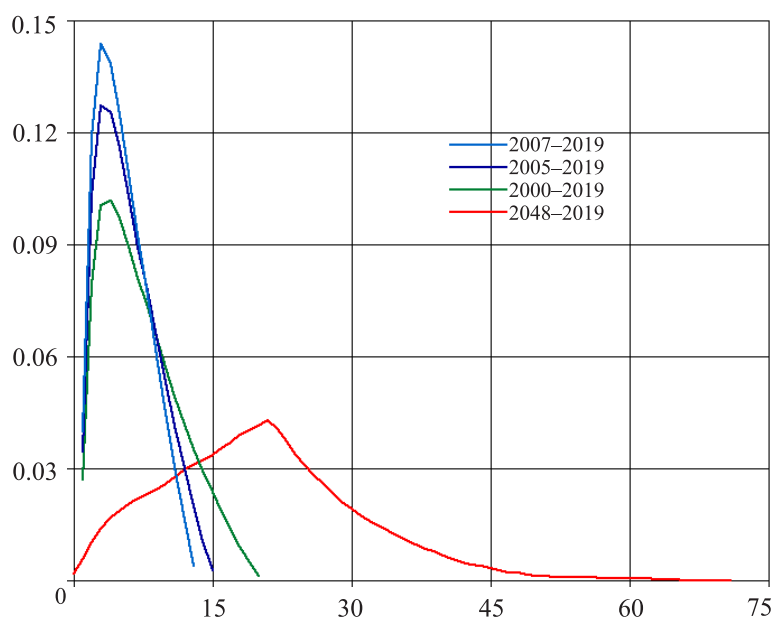
Распределение  $D(t)$  для разных временных интервалов публикации цитируемых статей представлено на рис. 6. На оси абсцисс указаны интервалы, прошедшие после публикации цитируемых статей до момента цитирования, а на оси ординат — доля цитирующих статей  $n_i/n$  (шкала логарифмическая). Для  $t_0 = 1968$  (также  $t_0 = 1988$ ) наблюдается линейный рост долей  $n_i/n$  по мере увеличения значения  $\theta_i$ . Для агрегированного периода  $t_0 = \{1948, 1949, \dots, 1997\}$  первые 21 год наблюдается линейное увеличение, а затем — экспоненциальный спад ( $\lambda = -0,132$ ). То есть цитирующие авторы забывают о публикациях быстрее, чем это сказывается на цитируемых публикациях. Как результат, распределение входящих степеней следует степенному закону.

Рис. 5. Распределение  $S(t)$ 

**Заключение.** Исследована сеть цитирования статей БД RePEc. Вычислительный эксперимент подтвердил влияние РА на скорость, с которой статьи получают цитирования от новых статей. Для узлов сети  $N_{REP}$  вероятность получения нового ребра в момент  $t$  пропорциональна входящей степени  $k$  и аппроксимируется линейной функцией. Как показывают теоретические исследования, именно линейность приводит к тому, что распределение входящих степеней узлов имеет хвост, следующий степенному закону. В данном случае параметр степенного закона  $\gamma = 2,89$  имеет значение, близкое к  $\gamma = 3$ , что характерно для БА-модели.

С одной стороны, согласно идеализированному РА, статьи с высокой степенью цитирования наиболее вероятно получают новые цитирования, с другой — экспоненциальный рост числа статей снижает шанс цитирования “старых” статей, причем основной рост числа цитирований приходится на первые два года после публикации. Для сети  $N_{REP}$  средний возраст цитирования составляет около 11 лет. Процесс старения может быть представлен экспоненциальной функцией  $f(t)$  на промежутке от 2–3 до 20–23 лет, прошедших после публикации. Заметим, что акт цитирования и дата публикации могут существенно отличаться друг от друга. Кроме того, процесс цитирования “молодых” публикаций не завершен. Эти факты могут существенно влиять на статистику. Еще одной особенностью сети цитирования является влияние содержания статей на цитируемость. Популярная в данное время тематика может привлечь цитирования. Утратившие внимание статьи также вносят диссонанс при рассмотрении профилей данных. Как результат — отсутствие монотонности в распределениях  $S(t)$  и  $D(t)$ .

Следует отметить, что данные ограничены цитированиями между отобранными статьями без учета других типов статей, а также статей вне базы, что также может исказить динамическую картину. Однако общие тенденции прослеживаются.

Рис. 6. Распределение  $D(t)$ 

## Список литературы

1. LOTKA A. J. The frequency distribution of scientific productivity // J. of the Washington Academy of Science. 1926, V. 16. P. 675–682.
2. SIMON H. A. On a class of skew distribution functions // Biometrika. 1955. V. 42. P. 425–440.
3. PRICE D. J. DE SOLLA. Networks of Scientific Papers // Science. 1965. V. 149. P. 510–515.
4. MERTON R. K. The Matthew Effect in Science // Science. 1968. V. 159, iss. 3810. P. 56–63.
5. COLE J. R., COLE S. Social stratification in science. Chicago, IL: University of Chicago. 1973.
6. PRICE D. J. DE SOLLA. A general theory of bibliometric and other cumulative advantage processes // J. of the American Society for Information Science. 1976. V. 27(5–5). P. 292–306.
7. TSALLIS C., DE ALBUQUERQUE M. P. Are citations of scientific papers a case of nonextensivity? // Eur. Phys. J. B. 2000. V. 13, iss. 4. P. 777–780.
8. REDNER S. How popular is your paper? An empirical study of the citation distribution // Eur. Phys. J. B. 1998. V. 4, iss. 2. P. 131–134.
9. PETERSON G. J., PRESSE S., DILL K. A. Nonuniversal power law scaling in the probability distribution of scientific citations // in Proc. Natl. Acad. Sci. USA. 2010. V. 107, iss. 37. P. 16023–16027.
10. BARABASI A.-L., ALBERT R. Emergence of scaling in random networks // Science. 1999. V. 286. P. 509–512.
11. REPEC. General principles. [Electron. Resource]. <http://repec.org/>.
12. KRAPIVISKY P. L., REDNER S., LEYVRAZ F. Connectivity of growing random networks // Phys. Rev. Lett. 2000. V. 85. P. 4629–4632.
13. ЩЕРБАКОВА Н. Г. Модели сетей с предпочтительным присоединением // Проблемы информатики. 2019. № 3. С. 46–61.
14. BARABASI A.-L., ALBERT R., JEONG H. Mean-field theory for scale-free random networks // Physica A. 1999. V. 272. P. 173–187.
15. DOROGOVTSSEV S. N., MENDES J. F. F., SAMUKHIN A. N. Structure of growing network with preferential linking // Phys. Rev. Lett. 2000. V. 85. P. 4633–4636.

16. BARABÁSI A. L., JEONG H., NEDA Z., RAVASZ E., SCHUBERT A., VICSEK T. Evolution of the social network of scientific collaborations // *Physica A*. 2002. V. 311. P. 590–614.
17. JEONG H., NEDA Z., BARABÁSI A. L. Measuring preferential attachment for evolving networks // *EuroPhysics Letters*. 2003. V. 61. P. 567–572.
18. Lehmann S., Lautrup B., Jackson A. D. Citation networks in high energy physics // *Phys. Rev. E*. 2003. V. 68, 026113.
19. REDNER S. Citation statistics from more than the century of physical review // *arXiv:physics/0407137*.
20. WANG M., YU G., YU D. Measuring the preferential attachment mechanism in citation networks // *Physica A*. 2008. V. 387. P. 4692–4698.
21. DOROGOVTSSEV S. N., MENDES J. F. F. Evolution of reference networks with aging // *Phys. Rev. E*. 2000. V. 62. P. 1842–1845.
22. ZHU H., WANG X., ZHU J-Y. Effect of aging on network structure // *Phys. Rev. E*. 2003. V. 68, 056121.
23. HAJRA K.B., SEN P. Modelling aging characteristics in citation networks // *Physica A*. 2006. V. 368. P. 575–582.
24. CLAUSET A., SHALIZI C. R., NEWMAN M. E. J. Power-law distributions in empirical data // *SIAM Review* V. 51. 2009. P. 661–703.
25. IGRAPH – The network analysis package [Electron. Resource]. <https://igraph.org/>.
26. POLLMANN T. Forgetting and the ageing of scientific publications // *Scientometrics*. 2000. V. 47, N 1. P. 43–54.
27. LEHMANN S., JACKSON A. D., LAUTRUP B. Life, death and preferential attachment // *EuroPhysics Letters*. 2005. V. 69. P. 298–303.



**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 автоматизации 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 rynek elektronnoj 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 го-

сударственного 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.



**Наталья Щербакова**

окончила Новосибирский государственный университет по специальности

„Математическая лингвистика“ в 1967 году. С 1967 г. работала в Институте математики СО РАН, затем в Институте автоматизации и электрометрии СО РАН в области создания программного обеспечения систем передачи данных. С 2000 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 — 16.12.2019*