УДК 577.214.626+316.452

ИССЛЕДОВАНИЕ СТРУКТУРЫ И ЭВОЛЮЦИИ СЕТЕЙ НАУЧНОГО СОАВТОРСТВА НА ОСНОВЕ АНАЛИЗА НОВОСИБИРСКИХ ПУБЛИКАЦИЙ В ОБЛАСТИ БИОЛОГИИ И МЕДИЦИНЫ

© 2014 г. **И.И. Титов[1, 2], А.А. Блинов[2]**

[1] Федеральное государственное бюджетное учреждение науки Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия, e-mail: titov@bionet.nsc.ru;
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

Поступила в редакцию 9 октября 2014 г. Принята к публикации 22 октября 2014 г.

Из-за разнообразия взаимодействий сообщества живой материи, от бактериальных колоний до человеческих сообществ, имманентно более сложны, нежели ансамбли частиц в неживой природе. Одним из проявлений внутри- и межгрупповых взаимодействий в социуме являются сети соавторства научных публикаций. В нашей работе рассмотрена такая сеть для новосибирского научного сообщества в области биологии и медицины. Используя базу данных PubMed, мы построили сеть и рассчитали ее статистические характеристики. Распределение организаций по научной активности оказывается распределением с толстым хвостом и подчиняется так называемому закону Парето: 83 % публикаций и 75 % авторов принадлежат примерно 20 % самых активным организациям. Сравнение сетей последних показывает, что сети вузов обладают более выраженным ядром, нежели сети научно-исследовательских институтов. Проведен анализ «демографической» структуры ныне активных авторов. Показано, что значительную долю составляют авторы с коротким публикационным стажем, а дефицит авторов наблюдается среди впервые опубликованных в 1991–1997 гг. В целом, динамика сети оказывается нестационарной с сохранением тенденции к повышению активности.

Ключевые слова: сеть соавторства, структура сети, эволюция сети, статистический анализ.

ВВЕДЕНИЕ

В наше время едва ли не самым важным условием производства научного знания является взаимодействие ученых. При непосредственном влиянии научных сотрудников друг на друга это взаимодействие реализуется в виде обмена мнениями, разделения труда и т. д., увеличивая продуктивность (Lee, Bozeman, 2005) и цитируемость (Sooryamoorthy, 2009; Gazni, Didegah, 2011), и фиксируется в виде совместного авторства публикации – явного продукта научного сотрудничества.

Данные о соавторстве позволяют формально построить упрощенную социальную сеть научных работников, в которой ее члены распространяют, интерпретируют и производят новое знание посредством социальных взаимодействий. Исследования последних 15 лет показали, что сети соавторства в разных областях науки организованы сходным образом: они кластеризованы и обладают малым диаметром (Newman, 2004).

Эти свойства не могут быть получены при случайно-равномерном размещении ребер и характерны также для других социальных сетей, таких как сети киноактеров и директоров компаний (Newman, 2003).

Особенно ярко социальные контакты научных работников должны проявляться в сообществах компактных, междисциплинарных и географически изолированных научных центров, например в Новосибирском научном

центре (ННЦ). Последнюю четверть века новосибирская, как и вся российская наука, не находилась в стабильных условиях, а пережила трансформацию общественного строя. Поэтому исследование структуры и эволюции сети соавторства Новосибирского научного центра представляет особый интерес.

МАТЕРИАЛЫ И МЕТОДЫ

Данные о научных публикациях были выделены из базы данных PubMed (состояние на август 2014 г.) фильтрацией по названию города. Всего была найдена 5 571 публикация. Отметим, что в базе PubMed присутствует лишь часть научных публикаций, а выбор PubMed из нескольких аналогичных баз данных обусловлен субъективно ее большей однородностью. Затем был составлен список организаций Новосибирска, которые присутствуют в аффилиации авторов публикаций, и добавлен ГНЦ «Вектор», который находится в п. г. т. Кольцово в непосредственной близости от Новосибирска.

Далее этот список был сокращен автоматической идентификацией вариантов названий одной и той же организации. На последнем этапе были вручную объединены те организации, которые меняли названия с сохранением юридического лица (Институт химической биологии и фундаментальной медицины Сибирского отделения Российской академии наук и др.). Окончательный список состоял из 62 организаций города Новосибирска, сотрудники которых имеют публикации в базе PubMed.

Для каждой публикации каждый из ее авторов отнесен к одной из его организаций. Если автор имел публикации с разными аффилиациями, он оказывался независимо включенным в соответствующие сети.

Далее для каждой организации список авторов сокращался автоматической идентификацией различия англоязычных написаний фамилий, а также отождествлением авторов без второго инициала. Окончательный список состоял из 8 162 авторов.

Статистический анализ сетей соавторства проведен при помощи пакета NetInference (Титов и др., 2013). Эволюция сети восстановлена при помощи определения соответствия между кластерами на соседних временных срезах.

РЕЗУЛЬТАТЫ

Сначала охарактеризуем публикационную активность в целом. Во-первых, все новосибирские институты СО РАН имеют публикации в базе PubMed, несмотря на ее специализацию в области биологии и медицины, – вероятно, вследствие присущей ННЦ интенсивности междисциплинарных контактов. Во-вторых, большое число организаций (часть из них не являются ни научными, ни учебными) представлены малым числом публикаций. В то же время более половины (51 %) публикаций принадлежат трем наиболее активным организациям. Поэтому неудивительно, что распределение числа организаций по числу публикаций выглядит как распределение с толстым хвостом (рис. 1). Интересно, что активность организаций подчиняется так называемому закону Парето: 12 (19,4 %) организациям соответствует 83 % статей и 75 % авторов. В дальнейшем мы сфокусировали свое внимание на рассмотрении именно этих 12 наиболее активных организаций: нескольких институтов СО РАН и СО РАМН, ГНЦ «Вектор» и двух вузов – НГУ и НГМУ.

Теперь обратимся к изменению публикационной активности новосибирской медико-биологической науки во времени. В целом можно отметить рост активности, однако очень неоднородный (рис. 2) – с отчетливыми пиками в 1991, 2003 и 2014 гг. Эта картина характерна как для менее активных, так и для большин-ства самых активных организаций, но наиболее ярко

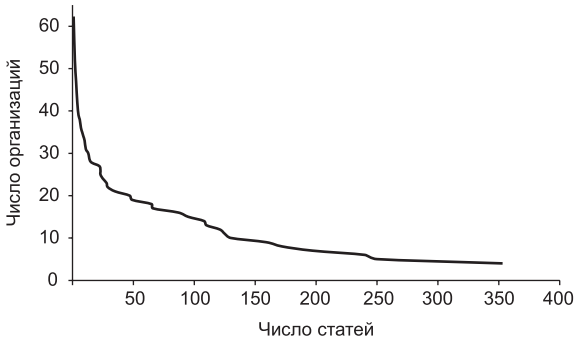

Рис. 1. Кумулятивное распределение числа новосибирских организаций по числу научных публикаций, аннотированных в базе PubMed. Функция $y(x)$ на графике показывает число организаций, которые имеют не менее x публикаций.

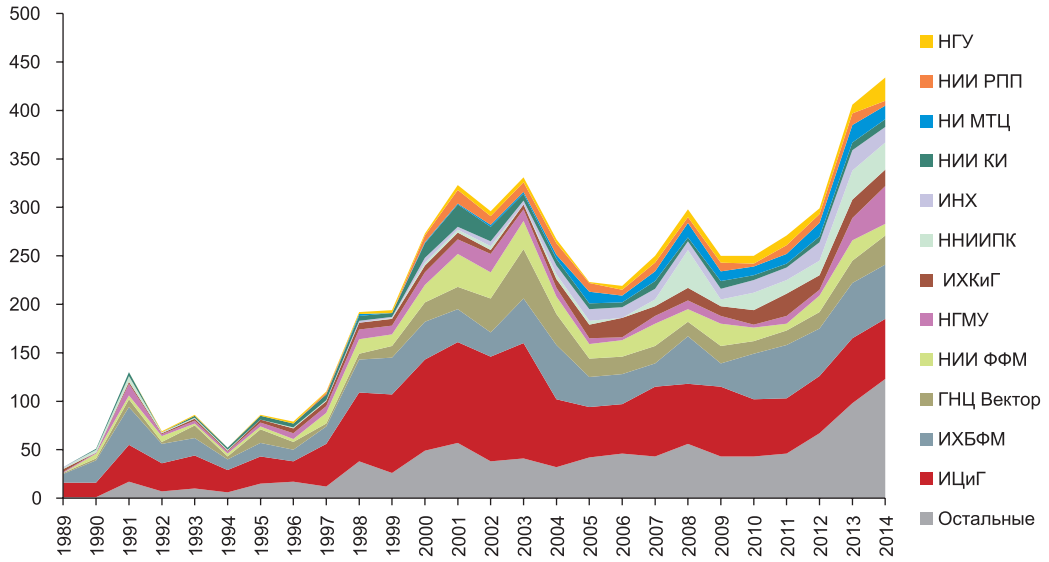

Рис. 2. Динамика числа публикаций в базе PubMed у 12 наиболее активных организаций Новосибирска. По оси абсцисс отложены года публикаций. Толщина полосы соответствующего цвета показывает число публикаций организации в заданный год. Таким образом, верхняя кривая отражает динамику числа публикаций всех новосибирских организаций. Организации представлены снизу вверх в порядке убывания полного числа публикаций в базе PubMed. Остальные 50 организаций выделены в группу «Остальные» (нижняя полоса). Напомним, что за 2014 г. данные неполные.

выражена у трех лидеров (ИЦиГ, ИХБФМ и ГНЦ «Вектор», рис. 2). Как показано ниже, яма 1990-х гг. и рост последних лет сопряжены с изменением числа научных работников, которые публикуются впервые. Исключение из общей тенденции составляет НГУ, число публикаций которого оставалось невысоким до тех пор, пока не выросло резко в 2014 г., вероятно, вследствие образования новых научных лабораторий.

В целом, оценка вкладов тех или иных факторов изменения научной активности требует отдельного исследования. Среди возможных факторов отметим изменение численности научных работников в результате внутренней и внешней эмиграции. В частности, снижение активности, наблюдаемое в середине 2000-х гг., могло быть вызвано прекращением новосибирской аффилиации эмигрантов.

Может ли нестационарный характер активности научных организаций быть связан с необычным поведением статистических характеристик авторов? Здесь под стационарностью процесса мы понимаем постоянство во времени его параметров, в том числе параметров роста. Для ответа на вопрос перейдем к статистическому описанию индивидуальной публикационной активности, в частности, проследим изменение активности авторов во времени. Сначала мы определили общее количество статей у каждого автора. В целом, распределение авторов по числу публикаций удовлетворяет степенному закону (данные не показаны), известному уже почти 90 лет (Lotka, 1926).

Далее для каждого автора, который имел публикацию в 2013–2014 гг., мы определили дату его первого появления в базе PubMed и построили распределение числа авторов по времени, которое прошло с момента первого появления в базе данных (рис. 3). Сходные распределения в виде пирамиды строят в обычных переменных для демографического анализа возрастной структуры населения. Молодыми учеными будем называть тех авторов, у которых публикационный стаж короче пяти лет. На полученной кривой можно выделить три участка: с начальным быстрым снижением, плато и яма (рис. 3).

Первый из участков, из сотрудников с публикационным стажем менее пяти лет, содержит две трети (66,4 %) из ныне активных научных работников. Этому участку можно сопоставить рост активности последних лет, где трехкратное

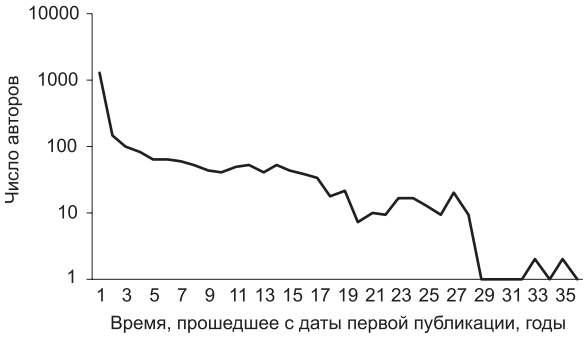

Рис. 3. Распределение числа авторов, которые опубликовали свои работы в 2013–2014 гг., по времени, прошедшему с даты их первой публикации (в полулогарифмических координатах).

увеличение числа авторов сопряжено с увеличением темпа публикаций более чем в полтора раза (см. рис. 2). Можно ли объяснить увеличение темпа публикаций ростом числа молодых ученых? Обе тенденции роста частично связаны друг с другом напрямую и частично вызваны действием третьего фактора.

Этот вывод обосновывается следующей простой оценкой. Для последнего четырехлетнего интервала времени мы определили доли публикаций, в которых присутствуют исключительно молодые или опытные научные работники.

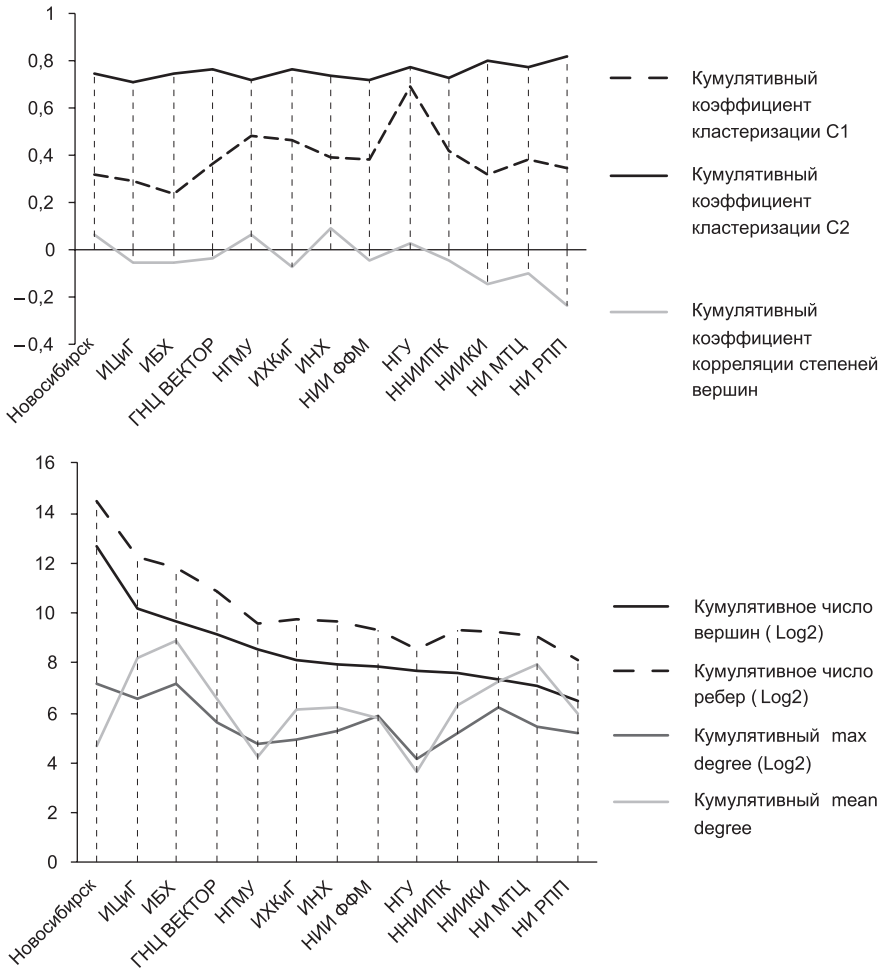

Рис. 4. Статистические характеристики сетей соавторства Новосибирска и 12 наиболее активных организаций. Некоторые характеристики стохастических сетей зависят от числа вершин в них, поэтому организации, в отличие от рис. 2, приведены в порядке уменьшения числа вершин в их сетях. Числа вершин и ребер, максимальная степень вершины показаны в полулогарифмических координатах. Для удобства общей шкалы по оси ординат приведены значения логарифмов этих характеристик.

Эти доли оказались равны 8 и 35 % соответственно, при двукратном перевесе численности первых. Тогда отношение продуктивности опытных и молодых научных работников можно оценить как $\frac{2 \cdot 35}{8} \approx 9$ раз. По той же логике продуктивность молодых научных работников при участии в работе опытных увеличивается в $\frac{100-35-8}{8} \approx 9$ раз. И наоборот, участие молодых научных работников в работе опытных повышает продуктивность работы последних в $\frac{100-35-8}{35} \approx 1{,}6$ раза.

Следует подчеркнуть крайнюю примитивность приведенных оценок, поскольку в них пренебрегается такими существенными обстоятельствами, как разное качество публикаций, индивидуальность авторов, неаддитивность и неравенство реальных вкладов соавторов, вариации численности и состава отдельных коллективов и обеих групп в целом, организация труда, сложившаяся структура научных коллективов и т. д.

Учет вышеназванных обстоятельств и накопленный к настоящему моменту объем публикаций позволят создавать точные наукометрические модели для прогноза эффективности работы научных организаций.

На втором участке число сотрудников слабо зависит от времени первой публикации вплоть до ямы между 17 и 22 годами. Это падение соответствует авторам, первые публикации которых появились в 1991–1997 гг., и может быть одним из факторов снижения общей активности, наблюдаемого в это же время (см. рис. 2).

От описания индивидуальной активности перейдем к рассмотрению взаимодействий авторов. Для этого мы построили сети соавторства, где каждому автору для каждой аффилиации соответствует своя вершина графа, а совместной публикации соответствует ребро между вершинами. Вес как вершины, так и ребра этого графа зависит от числа публикаций.

Взвешивание ребер и вершин графа может использоваться для количественной оценки социального влияния в сети, что было реализовано в нашей предыдущей работе для идентификации научных коллективов (Титов и др., 2013). Здесь же мы рассматриваем статистику построенных сетей соавторства (рис. 4).

Разнообразие значений характеристик сетей и, в частности, немонотонный характер их зависимости от числа вершин свидетельствуют об индивидуальности архитектуры рассмотренных сетей, поскольку их нельзя получить друг из друга масштабным преобразованием. Из рис. 4 видно, что во всех 12 сетях и сети новосибирского сообщества наблюдается высокая степень кластеризации.

Этот факт отмечен ранее для сетей соавторства разных областей знаний (Newman, 2004). При этом из общей выборки выделяются оба вуза: НГМУ и НГУ. Для них характерны высокие значения глобальных коэффициентов кластеризации, низкая плотность и ассортативность (взаимное притяжение вершин с близким числом входящих ребер).

Все эти свойства свидетельствуют об иной организации сетей вузов в сравнении с научными институтами: для первых характерно наличие единственного ярко выраженного ядра сети в сравнении со вторыми, более однородными и децентрализованными.

ЗАКЛЮЧЕНИЕ

Исследование сложных систем удобно сводить к изучению свойств сетей, которые моделируют эти системы. Для таких сетей часто характерны особая архитектура, богатая динамика и необычная эволюция. В этой работе мы исследуем сеть соавторства ННЦ, построенную по данным публикаций из базы PubMed. Мы показываем, что активность медико-биологического сообщества Новосибирска на протяжении последней четверти века, в целом демонстрируя значительный рост и обладая организацией типичного научного сообщества, испытавала драматические пики и провалы. Нестационарность эволюции сети проявляется во временных характеристиках авторов, в то время как структурная статистика авторов соответствует хорошо известному распределению активности.

БЛАГОДАРНОСТИ

Работа поддержана интеграционным проектом СО РАН № 21 и проектом фундаментальных исследований СО РАН VI.61.1.2.

ЛИТЕРАТУРА

Титов И.И., Блинов А.А., Рудниченко К.А., Крутов П.В., Казанцев А.Л., Куликов А.И. NETINFERENCE: набор программ для анализа структуры и динамики сетей // Вавиловский журнал генетики и селекции. 2013. Т. 17. № 4/1. С. 615–619.

Gazni A., Didegah F. Investigating different types of research collaboration and citation impact: a case study of Harvard University's publications // Scientometrics. 2011. V. 87. No. 2. P. 251–265.

Lee S., Bozeman B. The impact of research collaboration on scientific productivity // Social Studies of Science. 2005. V. 35. No. 5. P. 673–702.

Lotka A.J. The frequency distribution of scientific productivity // J. Wash. Acad. Sci. 1926. V. 16. No. 12. P. 317–324.

Newman M.E.J. The structure and functions of complex networks // SIAM review. 2003. V. 45. No. 2. P. 167–256.

Newman M.E.J. Coauthorship networks and patterns of scientific collaboration // PNAS. 2004. V. 101. No. S. 1. P. 5200–5205.

Sooryamoorthy R. Do types of collaboration change citation? Collaboration and citation patterns of South African science publications // Scientometrics. 2009. V. 81. No. 1. P. 177–193.

EXPLORING THE STRUCTURE AND EVOLUTION OF THE NOVOSIBIRSK BIOMEDICAL CO-AUTHORSHIP NETWORK

I.I. Titov[1, 2], A.A. Blinov[2]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia,
e-mail: titov@bionet.nsc.ru;
[2] Novosibirsk National Research State University, Novosibirsk, Russia

Summary

The interaction diversity within the communities of living matter, from bacterial colonies to human societies, makes them inherently more complex than ensembles of particles in inanimate nature. Co-authorship networks are a particular case of intra- and inter-group social interactions. In this paper, we analyze the Novosibirsk biomedical scientific community as an example of such a network. Using the PubMed database, we have built a community network and calculated its statistics. The distribution of organizations by scientific activity has a fat tail and obeys the Pareto principle: 83% of publications and 75% of authors belong to the 20% of the most active organizations. A comparison of their networks shows that networks of the universities have a more pronounced core rather than those of research institutions. We have plotted the "demographic" structure of currently active authors and found out two facts: (1) an abundance of authors with short "publication experience" and (2) a deficit of authors whose first publication is dated back to 1991-1997. In general, the network dynamics is non-steady, and the activity tends to increase.

Keywords: co-authorship network, network structure, network evolution, statistical analysis.