

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

MATHEMATICAL MODELING

Научная статья

УДК 004.93

doi: 10.17223/19988605/58/3

Сравнительный анализ моделей случайных графов

Владимир Никифорович Кузьмин¹, Федор Леонидович Шуваев², Максим Валерьевич Розганов³^{1, 2, 3} Военно-космическая академия им. А.Ф. Можайского, Санкт-Петербург, Россия¹ kuzmindvn@ya.ru² cadetfed@mail.ru³ mrozganov@mail.ru

Аннотация. Проведен сравнительный анализ основных математических моделей случайных графов. Рассмотрены экспоненциальные модели случайных графов, статистически доказана их адекватность реальным сетям. Показано, что такие модели расширяют инструментарий для исследования реальных сетей. Предложен усовершенствованный алгоритм настройки экспоненциальных моделей случайных графов.

Ключевые слова: граф; экспоненциальная модель; критерий согласия; мера близости, оценка

Для цитирования: Кузьмин В.Н., Шуваев Ф.Л., Розганов М.В. Сравнительный анализ моделей случайных графов // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2022. № 58. С. 23–34. doi: 10.17223/19988605/58/3

Original article

doi: 10.17223/19988605/58/3

Comparative analysis of random graph models

Vladimir N. Kuzmin¹, Fedor L. Shuvaev², Maxim V. Rozganov³^{1, 2, 3} Mozhaisky Military Space Academy, St. Petersburg, Russian Federation¹ kuzmindvn@ya.ru² cadetfed@mail.ru³ mrozganov@mail.ru

Abstract. The study of networks of diverse nature, which are citation networks, social networks, or information and telecommunications networks, is given attention in various fields of science: physics, biology, computer science, and mathematics. In-depth studies of the topological properties of such networks contribute to the understanding of its functionality and other inherent features, such as stability. In studies of a diverse nature, much attention is paid to the construction and verification of the adequacy of models, since the reliability of the results depends on it. In the field of network research based on graph theory, there are about ten mathematical models, the main ones being the Erdos-Renyi random graph model (ER), the freely scalable Barabashi-Albert model (BA), and the exponential random graph model (ERGM). The most important quality of the EMSG models is shown in the fact that there are 15 basic network statistics that characterize real networks and allow us to draw statistically significant conclusions about which statistics control the network under study.

In this study, a comparative analysis of graph models is carried out, and a statistical assessment of their adequacy to real networks is carried out based on the calculation of the Euclidean distance between the networks. To calculate the Euclidean distance, graph models were presented in the space of parameters of graph characteristics, such as: average path distance (SR), density (PG), assortativity (AS), degree centrality (DC), closeness centrality (CC), betweenness centrality (BC), eigen vector centrality (EVC), transitivity (Tr), diameter (Di). The parameter space was reduced on the basis of McKay statistics. These statistics showed that the characteristics of the AS, BC, and EVC did not meet the requirements of data uniformity, so they were excluded.

As a result of the study, it is shown that the closest to real networks that reflect the types of human activity are ERGM. This proximity to real networks was achieved, among other things, by an improved algorithm for configuring ERGM, which consists in selecting the best combination of terminals according to the Akaike information criterion.

Keywords: graph; exponential model; criterion of agreement; measure of proximity; assessment

For citation: Kuzmin, V.N., Shuvaev, F.L., Rozganov, M.V. (2022) Comparative analysis of random graph models. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitel'naja tehnika i informatika – Tomsk State University Journal of Control and Computer Science*. 58. pp. 23–34. doi: 10.17223/19988605/58/3

В настоящее время активно развивается научное направление анализа сетей, базирующееся на положениях теории графов. Это обусловлено существенным ростом вычислительных мощностей ЭВМ, появлением языков высокого уровня и специального программного обеспечения, позволяющего эффективно исследовать сети. При этом под сетями подразумеваются сети различной природы: сети взаимодействия белков в живых организмах, сети цитирования, информационно-коммуникационные сети (ИКС), социальные сети и др.

Для исследования сетей к настоящему времени разработан достаточно богатый инструментарий, базирующийся на графовых моделях [1–4]. С помощью таких моделей становится возможным тестировать гипотезы о сетевых структурах, процессах их формирования и динамике эволюции.

Кроме того, активное развитие методов машинного обучения и анализа данных открывает новые перспективы для использования моделей случайных графов, например при разработке систем распознавания топологий исследуемых сетей. Эти системы активно применяются для решения задач обеспечения информационной безопасности в информационно-телекоммуникационных сетях, при анализе социальных сетей и сетей цитирования. Формирование множества размеченных образов на основе той или иной модели случайного графа позволяет проверять точность систем распознавания.

Вместе с тем при всем многообразии моделей случайных графов не существует достаточных оснований для формирования суждения о том, какие же из этих моделей наиболее адекватно отражают природу реальных сетей. Целью исследования, основные результаты которого представлены в настоящей публикации, является разработка методической схемы сравнительного анализа адекватности моделей случайных графов реальным сетям.

1. Математические модели случайных графов

Современная теория графов предлагает около десяти математических моделей случайных графов, основными из которых считаются модель случайного графа Эрдеша–Реньи (ЭР), свободно масштабируемая модель Барабаши–Альберт (БА) и экспоненциальная модель случайного графа (ЭМСГ) [1, 3].

1.1. Модель случайного графа Эрдеша–Реньи

Модель случайного графа впервые предложена венгерскими математиками П. Эрдешем и Э. Реньи. Принцип построения случайного графа заключается в поочередном выборе натурального целого числа n из множества $\{n \in N\}$ до $n_{\text{треб}}$, таким образом $V = \{v_1, \dots, v_n\}$ – множество вершин моделируемого случайного графа. При этом $E = \{e_1, \dots, e_n\}$ – все возможные ребра, которые можно образовывать между парами элементов из V . Далее по схеме Бернулли выбираются ребра из множества $\{e_1, \dots, e_n\}$ с заданной вероятностью успеха p , и в случае успеха очередное ребро вносится в строящееся

ся множество ребер E , а в случае неудачи – не вносится. Итогом этой схемы является реализация случайного графа $G = \{V, E\}$ ЭР [3].

1.2. Модель предпочтительного присоединения Барабаши–Альберт

В реальных сетях новые вершины с большей вероятностью соединяются с вершинами, которые наиболее центральны по степени. Для анализа реальных сетей разработаны так называемые свободно масштабируемые модели, и самой распространенной среди них является модель предпочтительного присоединения, предложенная А. Барабаши и Р. Альберт [5, 6]. Эта модель описывает формирование случайного графа по следующим правилам:

1. В начальный момент времени $t = 0$ задается V_t несвязных вершин.
2. На каждом временном шаге $t = (1, 2, 3, \dots)$ добавляется новая вершина с E_t ребрами;
3. Количество ребер, с которыми приходит в граф новая вершина – фиксировано, но соединяется она с уже существующей вершиной сети с вероятностью, пропорциональной степени этой вершины.

Свойства реальных сетей наиболее полно описывает версия модели БА с фиксированным параметром распределения вероятностей соединения вершин [6] и вариативным числом новых ребер. Отличие заключается в том, что в модель вводится кортеж распределения вероятностей образования вершин, в котором p_0 – вероятность изолированности вершины, p_1 – вероятность соединения вершины с одной вершиной, p_2 – вероятность соединения вершины с двумя вершинами, p_3 – вероятность соединения вершины с тремя вершинами.

1.3. Экспоненциальная модель случайного графа

Экспоненциальные модели случайных графов нашли широкое применение при решении различных задач анализа сетей, в первую очередь социальных. При формировании ЭМСГ математический аппарат позволяет учитывать множество параметров, важных именно для рассматриваемой сети, например меры центральности [7–12], плотность графа, ассортативность, триадный вектор и т.п. [13–17].

Формируемая модель представляет собой граф G из пространства всех возможных графов для n вершин Λ . Ряд распределения графов по пространству Λ оценивается на основе следующего выражения:

$$P(G) = \frac{1}{c} \cdot e^{\Theta_1 \cdot z_1(G) + \Theta_2 \cdot z_2(G) + \dots + \Theta_p \cdot z_p(G)} \quad (1)$$

где c – нормирующая константа, $\sum_{G \in \Lambda} e^{\Theta_1 \cdot z_1(G) + \Theta_2 \cdot z_2(G) + \dots + \Theta_p \cdot z_p(G)}$; Θ_k – коэффициент, определяющий

склонность сети к образованию связей k -го типа (например, ребер, 2-star, 2-path и т.п.); $z_k(G)$ – сетевая статистика / терм (от англ. *term*), характеризующая число связей k -го типа в модели. Таким образом вероятность графа G определяется суммой сетевых статистик, взвешенных коэффициентами, в степени экспоненты.

Мощность множества Λ для ненаправленного графа равна 2^n , для направленного – $2^{n(n-1)}$. Не сложно заметить, что даже для небольших n поиск графа с максимальной вероятностью методом полного перебора является крайне ресурсоемким процессом. Один из методов формирования ЭМСГ, получивший широкое распространение на практике, представляет собой Марковские цепи по методу Монте-Карло (МЦМК; англ. MCMC – Markov chain Monte Carlo), в частности реализация модели по алгоритму Метрополиса–Гастингса [13, 14]. В ходе таких испытаний параметры модели подбираются под реальную сеть, в результате чего исследователь получает представление о лежащих в основе сети процессах.

ЭМСГ являются относительно новым средством анализа, поэтому рассмотрим содержание основных этапов моделирования.

Формирование ЭМСГ по алгоритму Метрополиса–Гастингса – это итеративный процесс последовательного выбора графов ($G_0, G_1, \dots, G_t, \dots, G_T$) из множества Λ , который включает в себя ряд этапов.

На первом этапе по реализации анализируемой сети оцениваются число вершин n и интересующие коэффициенты Θ_k .

На втором этапе задается начальный граф G_0 с n вершинами. Обычно начальный граф формируется без связей между вершинами, при этом он также может формироваться со случайными связями.

Этапы с третьего по пятый итеративно повторяются T раз. Необходимое число итераций для каждой сети определяется отдельно и зависит от скорости сходимости модели.

На третьем этапе случайным образом выбирается пара вершин h и k текущего графа G_t . Если между этими узлами связь отсутствует, то эта связь добавляется, если связь есть, то удаляется. Таким образом, формируется новый граф G_{t+1} .

На четвертом этапе, исходя из выражения (1), вычисляется отношение вероятностей графов: нового G_{t+1} и текущего G_t :

$$\frac{P(G_{t+1})}{P(G_t)} = \frac{e^{\Theta_1 \cdot z_1(G_{t+1}) + \Theta_2 \cdot z_2(G_{t+1}) + \dots + \Theta_p \cdot z_p(G_{t+1})}}{e^{\Theta_1 \cdot z_1(G_t) + \Theta_2 \cdot z_2(G_t) + \dots + \Theta_p \cdot z_p(G_t)}} = e^{\Theta_1 \cdot (z_1(G_{t+1}) - z_1(G_t)) + \Theta_2 \cdot (z_2(G_{t+1}) - z_2(G_t)) + \dots + \Theta_p \cdot (z_p(G_{t+1}) - z_p(G_t))}.$$

На пятом этапе принимается решение, какой граф выбирается в качестве текущего для следующей итерации. Для этого оценивается вероятность принятия нового графа G_{t+1} в качестве текущего:

$$P_{t+1} = \min \left\{ 1, \frac{P(G_{t+1})}{P(G_t)} \right\}.$$

Если $P(G_{t+1}) > P(G_t)$, то $P_{t+1} = 1$ и в качестве текущего графа однозначно выбирается G_{t+1} . Если $P(G_{t+1}) < P(G_t)$, то $P_{t+1} < 1$. В этом случае генерируется одна реализация случайной величины u с равномерным законом распределения в диапазоне от 0 до 1. Когда $u < P_{t+1}$, в качестве текущего графа выбирается G_{t+1} , когда $u \geq P_{t+1}$ – G_t .

В результате выполнения всех итераций формируется граф G_T , который представляет собой ЭМСГ анализируемой сети.

В исследованиях [15–17] показано, что существует около 15 основных базовых сетевых статистик, характеризующих реальные сети. Одно из важнейших качеств ЭМСГ заключается в том, что с их помощью можно сделать статистически значимые выводы: какие термы управляют исследуемой сетью. В табл. 1 приведены базовые термы, формирующие сети.

Таблица 1

Основные термы экспоненциальных моделей случайных графов

№ п/п	Наименование термина		Описание
	русскоязычное	англоязычное	
Общие			
1	Взаимосвязанность	Mutual	Взаимосвязанность сети
2	Эквивалентность ребер	Edges	Эквивалентность количества ребер
3	Изолированные вершины	Isolates	Эквивалентность количества изолированных вершин
4	Средняя степень вершины	Meandeg	Равенство средних степеней для вершин моделируемой сети
5	Плотность сети	Density	Равенство плотности сети
Структурные			
6	Транзитивность	Transitive	Эквивалентность транзитивности сети
7	Интранзитивность	Intransitive	Равенство триад типов 111D, 201, 111U, 021C и 030C
8	Двойной путь	Twopath	Эквивалентность количества элементов, представляющих собой минимально возможный путь от одной вершины к другой через транзитивную вершину
9	Симмелианская триада	Simmelian	Равенство триад simmelian (основаны на сильной взаимосвязи между двумя вершинами через третью)
10	Балансные триады	Balance	Равенство триад типов 102 и 300

Приведенные в табл. 1 термины предназначены только для направленных сетей. Вместе с русским наименованием указано и английское оригинальное имя термина; это сделано потому, что на русский язык они переводятся впервые. Таким образом, основная особенность ЭМСГ состоит в том, что при формировании графов, в частности при принятии решения о наличии связи между любыми двумя узлами, учитывается структура всего графа, а именно сколько интересующих базовых элементов в графе уже присутствует.

Термы разделены на две большие группы – общие и структурные. Общие отражают базовые характеристики графа, такие как степенное распределение, степень взаимосвязанности вершин графа, доля изолированных вершин и плотность графа. Структурные термины представляют собой минимальные элементы построения сети различной природы, иными словами, для каждой сети характерны свои уникальные термины [15–17].

Одним из преимуществ ЭМСГ является возможность работы с различными типами предикторов. Их можно разбить на четыре большие группы: предикторы узлов (ПУ), предикторы ребер (ПР), предикторы диад (ПД) и предикторы локальных структур (ПЛС). Существует более сотни вариантов предикторов, в настоящей статье, без ущерба для общности представления, ограничимся ПУ и ПР. Примерами ПУ являются номинативные или численные характеристики узла, например расстояние между узлами или гендерная принадлежность человека, представленного узлом, если рассматривается социальная сеть. Примером ПР является интенсивность передачи данных между узлами в ИКС. На этапе моделирования ЭМСГ оценивается статистическая значимость предикторов, и если гипотеза о их значимости подтверждается, то они включаются в модель, что может значительно повысить ее адекватность.

2. Подбор параметров математических моделей случайных графов

Исследование реальных сетей с использованием моделей случайных графов реализуется в два этапа. На первом этапе осуществляются выбор математической модели и подбор ее параметров, наиболее полно соответствующих исследуемой сети. На втором этапе проводится исследование сети с использованием выбранной модели. В этой связи особую важность имеет именно первый этап, представляющий собой этап структурной и параметрической идентификации сети, состоящий в отыскании наилучшей структуры и параметров модели.

Исходным параметром при моделировании является количество вершин n в исследуемой реальной сети. Это начальная установка, на основании которой осуществлялось моделирование в настоящем исследовании.

2.1. Подбор параметров для модели случайного графа Эрдеша–Реньи

Модель ЭР является двухпараметрической, т.е. при ее формировании используются число вершин n и число ребер m . Количество вершин фиксировано и соответствует исходной сети, а число ребер изменяется от 1 до $m_{\max}$, где $m_{\max}$ – максимально возможное количество ребер при фиксированном количестве вершин ($m_{\max} = n^2 - n$). Перебор параметров модели ЭР заключается в формировании множества реализаций случайных графов с заданным числом вершин для каждого значения варьируемого параметра – числа ребер.

2.2. Подбор параметров для модели предпочтительного присоединения Барабаши–Альберт

Как было отмечено, свойства реальных сетей наиболее полно описывает версия модели БА с кортежем распределений вероятностей образования вершин $P = \langle p_0, p_1, p_2, p_3 \rangle$. При этом, чтобы исключить наличие изолированных вершин, вероятность p_0 приравнивается нулю. Каждая из вероятностей образования вершин (кроме p_0) принимает значение на интервале (0; 1). В такой ситуации кортеж вероятностей всегда состоит из трех варьируемых элементов, а общее число кортежей равно $9^3 = 729$. Поиск параметров модели БА состоит в последовательном переборе возможных кортежей

вероятностей P с последующим формированием множества реализаций случайных графов для каждого кортежа.

2.3. Подбор параметров для экспоненциальной модели случайного графа

Для ЭМСГ оцениваемым параметром, позволяющим выбрать оптимальную модель, являются информационный критерий Акаике (AIC) и значение статистической значимости p -value для каждого термина. По AIC выбирается модель ЭМСГ с оптимальной комбинацией термов, одновременно с этим термины должны быть статистически значимы на заданном уровне p -value. Считается, что чем ниже значение AIC, тем лучше модель [18, 19]. Для подбора параметров ЭМСГ использован следующий алгоритм:

1. По 10 терминам, представленным в табл. 1, рассчитываются все возможные комбинации по 2, 3 и 4 термина;
2. Осуществляется построение ЭМСГ по каждой полученной комбинации из 2, 3, и 4 термов;
3. Выбираются только те модели, у которых статистическая значимость p -value термов не менее 0,05, а затем из них выбирается модель с наименьшим AIC.

3. Оценка адекватности модели

3.1. Структурная схема оценки адекватности модели

Под адекватностью модели понимают степень ее соответствия тому реальному явлению или объекту, для описания которого она строится. Вместе с тем создаваемая модель ориентирована, как правило, на исследование определенного подмножества свойств этого объекта. Поэтому можно считать, что адекватность модели определяется степенью ее соответствия не столько реальному объекту, сколько свойствам объекта в предметной области проводимого исследования.

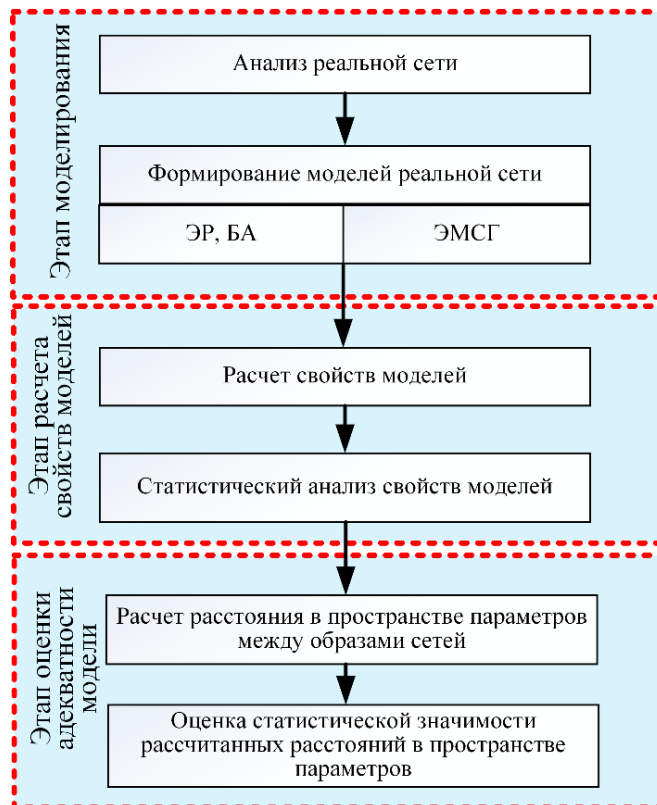


Рис. 1. Структурная схема сравнительного анализа адекватности моделей случайных графов

Fig. 1. Block diagram of the comparative analysis of mathematical models of random graphs

Если объектом исследования является сеть, то подмножеством ее свойств являются следующие характеристики, описанные в теории графов [7–12]: среднее расстояние пути (СРс), плотность (ПГ), ассортативность (АС), централизация по степени (СЦ), централизация по близости (СБ), централизация по посредничеству (СП), централизация по собственному вектору (СВЦ), транзитивность (Тр), диаметр (Ди) [13–17].

На рис. 1 представлена структурная схема проведения сравнительного анализа различных моделей реальных сетей. Она состоит из трех основных этапов, которые, в свою очередь, делятся на подэтапы.

Этап моделирования. На этом этапе производится первичный анализ реальной сети – объекта моделирования. В ходе анализа подсчитываются первичные характеристики реальной сети – число узлов и вершин, определяется ее тип – связная или несвязная, направленная или ненаправленная. На основе первичного анализа реальной сети формируются модели ЭР, БА и ЭМСГ.

Этап расчета свойств моделей. Для каждой реализации графа рассчитывается совокупность значений 9 основных свойств графов:

$$F = (f_{СРс}, f_{ПГ}, f_{АС}, f_{Ди}, f_{СЦ}, f_{СБ}, f_{СП}, f_{СВЦ}, f_{Тр}).$$

Элементы F формируют образ рассматриваемой модели графа или графа реальной сети в пространстве параметров. В результате расчета F для N реализаций графов получаем матрицу наблюдений $Y_{[N,n]} = \|f_{ij}\|_N^n$, где n – номер свойства графов.

Для полученной матрицы проводится статистический анализ изменчивости значений свойств графов, чтобы оценить однородность наблюдений. Статистический анализ осуществляется путем построения доверительных интервалов (ДИ) для коэффициента вариации ($кв$) [20]. В статистике принято считать, что совокупность наблюдений, коэффициент вариации которых превышает 33%, является неоднородной. Таким образом, из матрицы наблюдений исключаются признаки, которые выходят за границы рассчитанного для них ДИ, и те, $кв$ которых превышает 33%. ДИ (95%) рассчитываются на основе статистики Мак Кея [20, 21]:

$$\left(кв \left[\left(\frac{u_1}{v} - 1 \right) (кв)^2 + \frac{u_1}{v} \right]^{-1/2} \leq кв \leq кв \left[\left(\frac{u_2}{v} - 1 \right) (кв)^2 + \frac{u_2}{v} \right]^{-1/2} \right)$$

где $u_1 = \chi_{v,1-\alpha/2}^2$ и $u_2 = \chi_{v,\alpha/2}^2$, $100(1-\alpha/2)\%$ и $100(\alpha/2)\%$ – перцентили распределения Хи-квадрат с $v = n - 1$ степенями свободы [22].

Этап оценки адекватности модели. На этом этапе рассчитывается матрица евклидовых расстояний между модельными образами и образом исследуемой реальной сети в пространстве значений F . По расстоянию между образами моделей различных типов и реальной сетью осуществляется оценка статистической значимости различия между образом реальной сети и образом модели графа по критерию согласия Пирсона (Хи-квадрат). Считается, что те модели, которые статистически незначимо отличаются по расстоянию в пространстве параметров от исходной реальной сети, являются адекватными.

3.2. Результаты экспериментальных исследований

Проведем сравнительный анализ моделей ЭР, БА и ЭМСГ с реальной сетью отношений между менеджерами одной из компаний США (англоязычное название сети `ht_friends`), представленной в пакете «нетворкдата» (англ. `networkdata`), который является самым крупным набором сетевых данных для языка статистического моделирования R. Рассматриваемая реальная сеть № 1 является связной и направленной, имеет 21 вершину и 102 ребра.

Произведем моделирование путем формирования множеств реализаций случайных графов с использованием рассмотренных моделей на множестве их параметров, как показано в п. 2. Для каждой полученной модели рассчитаем образ F . В результате получены три матрицы наблюдений:

для БА размерности $Y_{[729000,9]}^{БА}$, для ЭР размерности $Y_{[420000,9]}^{ЭР}$. Размерности матриц обусловлены особенностями установки параметров моделей, указанных в п. 2. Отдельно производится расчет ЭМСГ, на основе его получено минимальное значение $AIC = 422$ при комбинации термов intransitive + mutual + meandeg + simmelianties.

Для полученных матриц наблюдений был осуществлен статистический анализ изменчивости, который заключался в измерении kv и построении ДИ для него. Результаты анализа представлены в табл. 2.

Таблица 2

Результаты расчета коэффициента вариации и ДИ

ДИ/Ф	СРс	ПГ	Ас	Ди	СЦ	СБ	СП	СВЦ	Тр
kv	21,6	31,6	46,5	30,3	29,2	103,3	10,7	41,3	27,6
ДИ нижний	20,7	29,0	44,3	29,0	28,0	96,3	9,3	39,4	26,2
ДИ верхний	22,5	33,5	48,9	31,8	30,6	111,9	11,4	43,4	28,3

Из табл. 2 видно, что однородны совокупности наблюдений СРс, Ди, СЦ, СП, ПГ и Тр (выделены серым), поэтому из дальнейшего анализа исключаем остальные три свойства и рассчитываем евклидово расстояние между образами сетей в пространстве шести оставшихся свойств графов.

На рис. 2 представлены зависимости евклидова расстояния между образами и тестовой сетью от номера модели для ЭР (черная пунктирная линия) и БА (красная сплошная линия), причем для наглядности представлены зависимости не только для сети ht_friends, но и для сетей № 2–4 из табл. 3. На рис. 2, а показано евклидово расстояние образов моделируемых сетей от образа тестовой сети № 1 ht_friends, на рис. 2, б – от образа сети № 2, на рис. 2, в – от образа сети № 3, на рис. 2, г – от образа сети № 4. Видно, что зависимость для модели ЭР изменчива, а для БА – однородна.

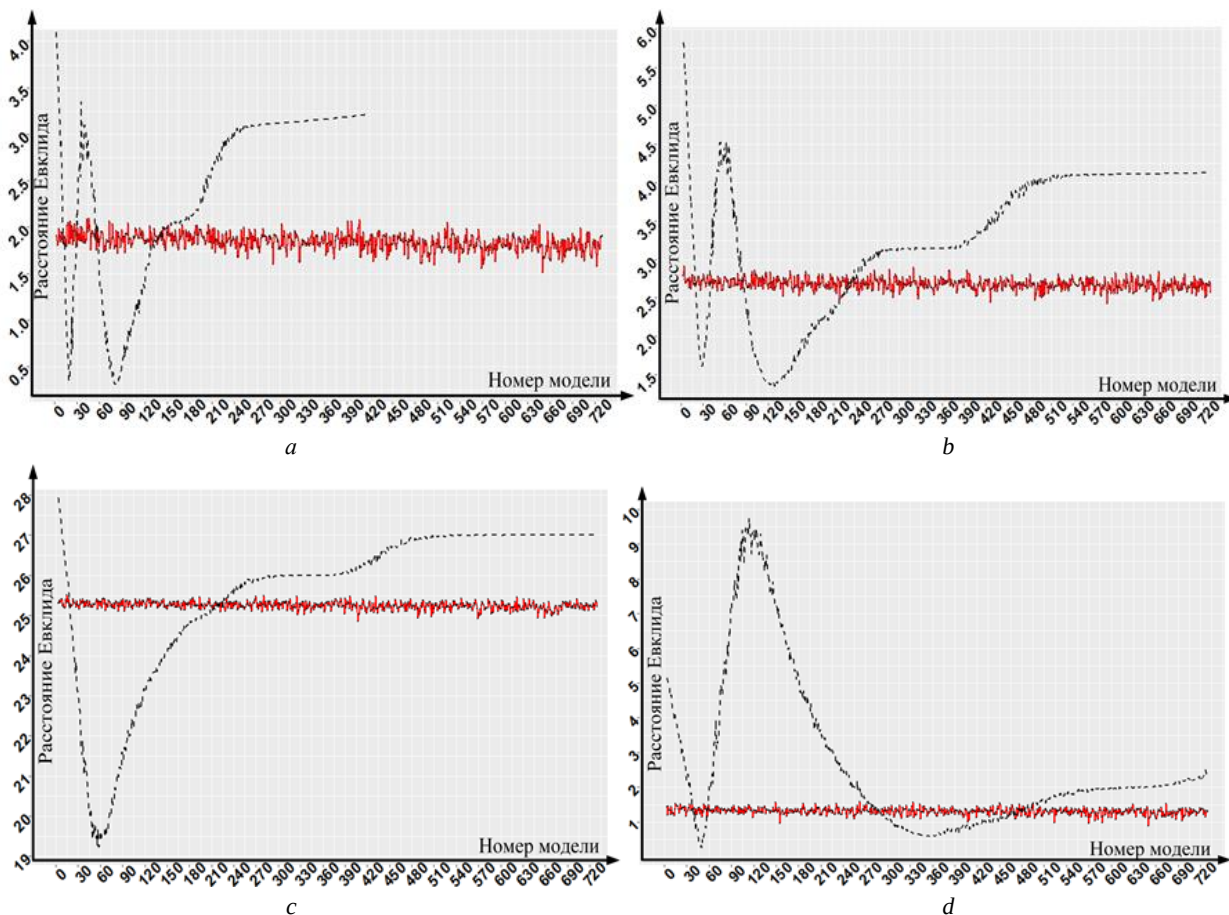


Рис. 2. Зависимости евклидова расстояния между образами сетей и номерами моделей сети
Fig. 2. Dependencies of Euclidean distance between network patterns and network model numbers

Из 729 моделей БА и 420 ЭР выбирается по одной, наиболее близкой к реальной сети. Для наилучшей по критерию AIC ЭМСГ аналогично рассчитываются образ и расстояние от него до образа реальной сети. По критерию согласия Пирсона (Хи-квадрат) осуществлена оценка статистической значимости расстояний между образами моделей и образом реальной сети. В соответствии с этой оценкой на уровне значимости $p\text{-value} = 0,005$ статистически значимыми являются различия в расстоянии между образом реальной сети и образами моделей ЭМСГ, ЭР, БА. Следовательно, не представляется возможным сформировать адекватную модель сети № 1, т.е. статистически незначимо отличающуюся от нее в пространстве параметров свойств графов.

3.3. Сравнительный анализ моделей графов по тестовому набору реальных сетей

Для обобщения полученного выше результата верификации моделей по представленной на рис. 1 схеме осуществлен сравнительный анализ рассмотренных моделей случайных графов с использованием данных о реальных сетях, представленных в пакете «нетворкдата».

Выбранные реальные данные отражают различные стороны человеческой деятельности: сеть № 1 – сеть дружеских взаимоотношений в кампусе Австралийского национального университета, № 2, 3 – ИКС в датацентрах одной из коммерческих организаций, № 4 – иерархия в стае птиц, № 5 – сеть взаимоотношений в колледже штата Иллинойс, № 6 – сеть менеджеров крупной корпорации, № 7 – рабочая ИКС в адвокатской конторе, № 8 – транспортная сеть авиаперевозок в США, № 9 – сеть межбелковых взаимодействий в человеческом организме. Результаты анализа представлены в табл. 3.

Таблица 3

Результаты сравнительного анализа моделей графов

Сеть	ЭМСГ			ЭР			БА		
	1	2	3	1	2	3	1	2	3
№ 1	intransitive + mutual + meandeg + simmelianties (AIC = 422)	5,05	3	$E = 58$	19,22	3	$P = \langle 0; 0,6; 0,2; 0,2 \rangle$	24,85	3
№ 2	twopath + density + intransitive (AIC = 630)	1,56	3	$E = 29$	0,41	Н	$P = \langle 0; 0,4; 0,2; 0,2 \rangle$	1,94	3
№ 3	intransitive + mutual + edges + simmelianties (AIC = 220)	0,48	Н	$E = 29$	0,46	Н	$P = \langle 0; 0,9; 0,1; 0,2 \rangle$	1,82	3
№ 4	intransitive + simmelianties + density + mutual (AIC = 689)	1,46	3	$E = 47$	0,69	3	$P = \langle 0; 0,8; 0,2; 0,2 \rangle$	1,11	3
№ 5	balance + intransitive + mutual + transitive (AIC = 3541)	14,14	3	$E = 102$	5,48	3	$P = \langle 0; 0,9; 0,1; 0,3 \rangle$	14,11	3
№ 6	density + edges + meandeg + mutual (AIC = 992)	0,1	Н	$E = 229$	0,12	Н	$P = \langle 0; 0,3; 0,1; 0,9 \rangle$	1,55	3
№ 7	intransitive + simmelianties + ПУ (AIC = 630,622)	0,42/0,23	Н	$E = 144$	0,45	3	$P = \langle 0; 0,7; 0,3; 0,2 \rangle$	0,69	3
№ 8	edges + mutual + simmelianties + 2ПУ (AIC = 482,432,411)	0,40/0,32/0,18	Н	$E = 147$	0,47	3	$P = \langle 0; 0,1; 0,2; 0,9 \rangle$	0,58	3
№ 9	edges + simmelianties + mutual + ПУ (AIC = 403)	0,32	Н	$E = 434$	0,60	3	$P = \langle 0; 0,8; 0,1; 0,2 \rangle$	10,45	3

Для каждого типа моделей созданы по три графы значений: графа 1 – уникальные параметры модели, графа 2 содержит евклидово расстояние в пространстве значений свойств между образами сформированной модели и реальной сети, графа 3 содержит индикатор значимости различия расстояния между образом модели и образом реальной сети. В табл. 3 представлено по одному представителю каждого типа моделей, которые в пространстве параметров свойств графов находятся ближе всего к сети-прототипу.

Реальные сети № 1–6 не содержали в себе атрибутов узлов и ребер, что не позволяло применять в ЭМСГ предикторы. Сети № 7–9 содержат атрибуты узлов, поэтому помимо термов при формировании ЭМСГ использовались ПУ, что заметно повысило адекватность модели: ЭМСГ – единственные модели, которые статистически незначимо отличаются от реальной сети. В графе 1 ЭМСГ наборы термов различны, что определяется механизмами установления связей между узлами в силу различной природы рассматриваемых сетей.

Необычными результатами отмечаются модели ЭР. Так, для сетей № 2, 3, 6 они показывают наилучшие значения по близости к реальной сети. Данное обстоятельство на данный момент не имеет научного объяснения и является направлением для дальнейших исследований.

Модели БА неожиданно отмечаются самыми низкими показателями. Из 9 моделей ни одна не показала статистически значимую близость к реальной сети. Можно отметить низкую изменчивость расстояния от реальной сети при моделировании БА, как видно из рис. 2.

Заключение

В результате анализа моделей случайных графов показано, что наиболее близкими к реальным сетям, отражающим виды человеческой деятельности, являются сети типа ЭМСГ и ЭР. Проведено имитационное моделирование, в ходе которого получены практические результаты, подтверждающие выводы анализа.

Предложенная схема анализа может применяться на практике для создания адекватных моделей реальных сетей. Исследование адекватности проведено с использованием оригинальной авторской методики, в основе которой лежит апробированный подход сравнения евклидова расстояния между образами сетей, рассчитанных на основе измерения основных свойств графов. При этом расстояние оценивается статистически на основе критерия согласия Пирсона.

Предложенный алгоритм настройки ЭМСГ, заключающийся в подборе наилучшей комбинации термов по информационному критерию Акаике, позволил повысить точность настройки модели и, следовательно, повысить ее сходство с реальными сетями.

Список источников

1. Chen P.-Y., Choudhury S., Hero A. Multi-centrality graph spectral decompositions and their application to cyber intrusion detection // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2016. P. 4553–4557.
2. Шуваев Ф.Л., Татарка М.В. Анализ динамики мер центральности математических моделей случайных графов // Научно-технический вестник информационных технологий, механики и оптики. 2020. Т. 20, № 2. С. 249–256.
3. Hartmann A., Mézard M. Distribution of diameters for Erdős-Rényi random graphs // Phys. Rev. 2018. V. 97, № 3. DOI: 10.1103/PhysRevE.97.032128
4. Gibson H., Vickers P. Using adjacency matrices to lay out larger small-world networks // Applied Soft Computing. 2016. V. 42. P. 80–92.
5. Шуваев Ф.Л., Татарка М.В. Анализ математических моделей случайных графов, применяемых в имитационном моделировании информационно-коммуникационных сетей // Вестник Санкт-Петербургского университета ГПС МЧС России. 2020. № 2. С. 67–77.
6. Barabasi A. Network Science. Glasgow : Cambridge university press, 2016. 453 p.
7. Bonchi F., De Francisci G., Riondato M. Centrality Measures on Big Graphs: Exact, Approximated, and Distributed Algorithms // Proc. of the 25th International Conference Companion on World Wide Web. 2016. P. 1017–1020.
8. Brandes U., Borgatti S., Freeman L. Maintaining the duality of closeness and betweenness centrality // Social Networks. 2016. V. 44. P. 153–159.
9. Minoo A., Salehzadeh-Yazdi A., Razaghi-Moghadam Z., Hennig H., Wolkenhauer O. A Systematic survey of centrality measures for protein-protein interaction networks // Systems Biology. 2018. V. 12 (1). P. 80–88.
10. Seo H., Thorson S.J. ERGM approach to press freedom, regime type, and Internet connectedness // First Monday. 2019. V. 24, № 9. URL: <https://firstmonday.org/ojs/index.php/fm/article/view/9428> (accessed: 05.09.2020).
11. Wang P., Robins G., Pattison P., Lazega E. Social selection models for multilevel networks // Social Networks. 2015. V. 35. P. 96–115.
12. Golshid Sharifnia S., Saghaei A. A statistical approach for social network change detection: an ERGM based framework // Communications in Statistics – Theory and Methods. 2020. DOI: 10.1080/03610926.2020.1772981

13. Marnissi Y., Chouzenoux E., Benazza-Benyahia A., Pesquet J. Majorize-minimize adapted Metropolis-Hastings algorithm // *IEEE Xplore*, 2020. P. 2356–2369.
14. Teixeira J., Stutz L., Knupp D., Silva Neto A. A new adaptive approach of the Metropolis-Hastings algorithm applied to structural damage identification using time domain data // *Applied Mathematical Modelling*. 2020. V. 82 (1). P. 587–606.
15. Zvereva O. Triad Census Usage for Communication Network Analysis // *CEUR Workshop Proc.* 2016. V. 1710. P. 378–389.
16. Brunson J.C. Triadic analysis of affiliation networks // *Network Science*. 2015. V. 3(4). P. 480–508.
17. Faust K. A puzzle concerning triads in social networks: Graph constraints and the triad census // *Social Networks*. 2010. № 32. P. 221–233.
18. Package ‘network’. URL: <https://cran.r-project.org/web/packages/network/network.pdf> (accessed: 05.09.2020).
19. Package ‘ergm’. URL: <https://cran.r-project.org/web/packages/ergm/ergm.pdf> (accessed: 05.09.2020).
20. Panichkitkosolkul W. Confidence intervals for the coefficient of variation in a normal distribution with a known population mean // *Journal of probability and statistic*. 2013. № 2. P. 34–46.
21. Albatineh A., Golam Kibria B., Wilcox M., Zogheib B. Confidence interval estimation for the population coefficient of variation using ranked set sampling: a simulation study // *Journal of applied statistics*. 2014. V. 41. P. 733–751.
22. Beigy M. Coefficient of Variation: cv_versatile. 2019. URL: https://cran.r-project.org/web/packages/cvcqv/vignettes/cv_versatile.html (accessed: 04.12.2020).

References

1. Chen, P.-Y., Choudhury, S. & Hero, A. (2016) Multi-centrality graph spectral decompositions and their application to cyber intrusion detection. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 4553–4557. DOI: 10.1109/ICASSP.2016.7472539
2. Shuvaev, F.L. & Tatarka, M.V. (2020) Dynamics of centrality measures of random graph mathematical models. *Nauchno-tehnicheskiiy vestnik informatsionnykh tekhnologiy, mekhaniki i optiki – Scientific and Technical Journal of Information Technologies, Mechanics and Optics*. 20(2). pp. 249–256. DOI: 10.17586/2226-1494-2020-20-2-249-256
3. Hartmann, A. & Mézard, M. (2018) Distribution of diameters for Erdős-Rényi random graphs. *Physical Review*. 97(3). DOI: 10.1103/PhysRevE.97.032128
4. Gibson, H. & Vickers, P. (2016) Using adjacency matrices to lay out larger small-world networks. *Applied Soft Computing*. 42. pp. 80–92. DOI: 10.1016/j.asoc.2016.01.036
5. Shuvaev, F.L. & Tatarka, M.V. (2020) Analysis of mathematical models of random graphs used in the simulation of information and communication networks. *Vestnik Sankt-Peterburgskogo universiteta GPS MChS Rossii – Bulletin of St. Petersburg University State Fire Service of the Ministry of Emergencies of Russia*. 2. pp. 67–77.
6. Barabasi, A. (2016) *Network Science*. Glasgow: Cambridge University Press.
7. Bonchi, F., De Francisci, G. & Riondato, M. (2016) Centrality Measures on Big Graphs: Exact, Approximated, and Distributed Algorithms. *Proceedings of the 25th International Conference Companion on World Wide Web*. pp. 1017–1020.
8. Brandes, U., Borgatti, S. & Freeman, L. (2016) Maintaining the duality of closeness and betweenness centrality. *Social Networks*. 44. pp. 153–159. DOI: 10.1016/j.socnet.2015.08.003
9. Minoo, A., Salehzadeh-Yazdi, A., Razaghi-Moghadam, Z., Hennig, H. & Wolkenhauer, O. (2018) A Systematic Survey of Centrality Measures for Protein-Protein Interaction Networks. *Systems Biology*. 12(1). pp. 80–88. DOI: 10.1186/s12918-018-0598-2
10. Seo, H. & Thorson, S.J. (2019) ERGM approach to press freedom, regime type, and Internet connectedness. *First Monday*. 24(9). [Online] Available from: <https://firstmonday.org/ojs/index.php/fm/article/view/9428> (Accessed: 4th December 2020).
11. Wang, P., Robins, G., Pattison, P. & Lazega, E. (2015) Social selection models for multilevel networks. *Social Networks*. 35. pp. 96–115. DOI: 10.1016/j.socnet.2014.12.003
12. Golshid, S.S. & Saghaei, A. (2020) A statistical approach for social network change detection: an ERGM based framework. *Communications in Statistics – Theory and Methods*. pp. 2259–2280. DOI:10.1080/03610926.2020.1772981
13. Marnissi, Y., Chouzenoux, E., Benazza-Benyahia, A. & Pesquet, J. (2020) Majorize-minimize adapted Metropolis-Hastings algorithm. *IEEE Xplore*. pp. 2356–2369. DOI: 10.1109/TSP.2020.2983150
14. Teixeira, J., Stutz, L., Knupp, D. & Silva Neto, A. (2020) A new adaptive approach of the Metropolis-Hastings algorithm applied to structural damage identification using time domain data. *Applied Mathematical Modelling*. 82. pp. 587–606. DOI: doi.org/10.1016/j.apm.2020.01.021
15. Zvereva, O. (2016) Triad Census Usage for Communication Network Analysis. *CEUR Workshop Proc.* 2016. V. 1710. pp. 378–389.
16. Brunson, J.C. (2015) Triadic analysis of affiliation networks. *Network Science*. 3(4). pp. 480–508.
17. Faust, K. (2010) A puzzle concerning triads in social networks: Graph constraints and the triad census. *Social Networks*. 32. pp. 221–233. DOI: 10.1016/j.socnet.2010.03.004
18. Cran.r-project.org. (n.d.) *Package ‘network’*. [Online] Available from: <https://cran.r-project.org/web/packages/network/network.pdf> (Accessed: 5th September).
19. Cran.r-project.org. (n.d.) *Package ‘ergm’*. [Online] Available from: <https://cran.r-project.org/web/packages/ergm/ergm.pdf> (Accessed: 5th September 2020).
20. Panichkitkosolkul, W. (2013) Confidence intervals for the coefficient of variation in a normal distribution with a known population mean. *Journal of Probability and Statistics*. pp. 34–46. DOI: 10.1155/2013/324940

21. Albatineh, A., Golam, K.B., Wilcox, M. & Zogheib, B. (2013) Confidence interval estimation for the population coefficient of variation using ranked set sampling: a simulation study. *Journal of Applied Statistics*. 41. pp. 733–751. DOI: 10.1080/02664763.2013.847405
22. Beigy, M. (2019) *Coefficient of Variation: cv_versatile*. [Online] Available from: https://cran.r-project.org/web/packages/cvcqv/vignettes/cv_versatile.html (Accessed: 4th December 2020).

Информация об авторах:

Кузьмин Владимир Никифорович – доктор военных наук, профессор, заслуженный деятель науки Российской Федерации, ведущий научный сотрудник Военно-космической академии им. А.Ф. Можайского (Санкт-Петербург, Россия). E-mail: kuzmindvn@ya.ru

Шуваев Федор Леонидович – кандидат технических наук, старший научный сотрудник Военно-космической академии им. А.Ф. Можайского (Санкт-Петербург, Россия). E-mail: cadetfed@mail.ru

Розганов Максим Валерьевич – старший научный сотрудник Военно-космической академии им. А.Ф. Можайского (Санкт-Петербург, Россия). E-mail: mrozganov@mail.ru

Вклад авторов: все авторы сделали эквивалентный вклад в подготовку публикации. Авторы заявляют об отсутствии конфликта интересов.

Information about the authors:

Kuzmin Vladimir Nikiforovich – Doctor of Science, Professor, Honored scientist of the Russian Federation, leading researcher, Mozhaisky Military Space Academy (St. Petersburg, Russian Federation). E-mail: kuzmindvn@ya.ru

Shuvaev Fedor Leonidovich – Candidate of Technical Science, Saint Researcher, Mozhaisky Military Space Academy (St. Petersburg, Russian Federation). E-mail: cadetfed@mail.ru

Rozganov Maxim Valeryevich – Saint Researcher, Mozhaisky Military Space Academy (St. Petersburg, Russian Federation). E-mail: mrozganov@mail.ru

Contribution of the authors: the authors contributed equally to this article. The authors declare no conflicts of interests.

Поступила в редакцию 20.04.2021; принята к публикации 28.02.2022

Received 20.04.2021; accepted for publication 28.02.2022