

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**

Кафедра математической кибернетики и компьютерных наук

**ИССЛЕДОВАНИЕ СЛУЧАЙНЫХ ВЕЛИЧИН, СВЯЗАННЫХ С
МОДЕЛЮ СЛУЧАЙНОГО ГРАФА БАРАБАШИ-АЛЬБЕРТ**

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 451 группы
направления 09.03.04 — Программная инженерия
факультета КНиИТ
Кадомцева Дмитрия Николаевича

Научный руководитель
зав. кафедрой, к. ф.-м. н., доцент _____

С. В. Миронов

Заведующий кафедрой
к. ф.-м. н., доцент _____

С. В. Миронов

СОДЕРЖАНИЕ

ВВЕДЕНИЕ	3
1 Постановка задачи	4
2 Исследование $M(k_i^n)$	6
2.1 Нахождение $M(k_i^2)$ и $D(k_i)$	6
2.2 Вывод формулы для $M(k_i^n)$	7
2.3 Нахождение $M(k_i^3)$, $M(k_i^4)$, а также коэффициента асимметрии	8
3 Исследование $M(s_i^2)$	10
3.1 Нахождение точного вида $M(s_i)$	10
3.2 Составление уравнения для поиска $M(s_i^2)$	11
3.3 Нахождение $M(k_i \cdot s_i)$	12
3.4 Нахождение $M(s_i^2)$ и $D(s_i)$	12
3.5 Проблемы с нахождением $M(s_i)$ для произвольного m	13
ЗАКЛЮЧЕНИЕ	15

ВВЕДЕНИЕ

Многие современные графы, такие как веб-графы или графы социальных связей, попадают в класс безмасштабных сетей, что означает, что они имеют степенное распределение степеней вершин, в то время как модели случайных графов, такие как модель Эрдеша-Реньи или модель Уоттса-Строгаца в этот класс не попадают. Однако модель Барабаши-Альберт попадает в этот класс и поэтому становится достаточно интересна в изучении, поскольку с помощью нее получается генерировать сети, сильно приближенные к реальным графам.

Целью данной работы является изучение таких случайных величин, связанных с моделью Барабаши-Альберт, как n -ая степень количества соседей вершины в момент времени t , что в свою очередь позволит найти такие характеристики, как дисперсия степени вершины в момент времени t , коэффициент асимметрии, а также дисперсия суммарной степени соседей вершины в момент времени t .

В результате написания работы должны быть решены следующие задачи:

- написание алгоритма для генерации графа по модели Барабаши-Альберт;
- моделирование таких величин, как n -ая степень количества соседей выбранной вершины, дисперсия количества соседей выбранной вершины, дисперсия суммарной степени соседей в момент времени t и т.д.;
- теоретическое обоснование полученных результатов;
- эмпирическое исследование поведения величин, для которых не удалось вывести явного закона.

1 Постановка задачи

С начала 21 века появилось много работ, посвященных изучению Интернета и моделированию его свойств с помощью случайных графов. Было показано, что уже существующая на тот момент модель Эрдёша-Реньи не подходит для данной задачи, поскольку сеть, сгенерированная по этой модели, и в отличие от Интернет-сети не является безмасштабной, то есть ее степенное распределение нельзя описать степенным законом. Однако еще задолго до выхода этой модели существовало такое понятие как предпочтительное соединение, и было показано, что случайные сети, которые строятся с использованием данной идеи, уже являются безмасштабными. Простыми словами предпочтительное соединение — это эвристика, по которой каждая новая вершина случайного графа соединяется с вершинами, имеющими наибольшую степень.

В 1999 году была предложена модель случайного графа, которая полностью основывается на линейном предпочтительном соединении, то есть когда вероятность линейно зависит от степени, и такая модель была названа в честь ученых, описавших ее — модель Барабаши-Альберт. В этой модели вероятность присоединения новой вершины к уже существующей i в момент времени $t + 1$ равняется

$$\frac{k_i}{\sum_{j=1}^t k_j} = \frac{k_i}{2mt}.$$

Процесс создания сети по этой модели можно описать так:

1. имеется какой-то начальный небольшой граф G ;
2. в течение каждой итерации в граф добавляется новая вершина, которая присоединяется к m уже существующим вершинам с помощью предпочтительного соединения.

Несмотря на кажущуюся простоту, модель Барабаши-Альберт достаточно хорошо способна моделировать реальные WWW сети и поэтому достойна изучения. В данной работе мы рассмотрим модель Барабаши-Альберт с алгоритмической точки зрения, попробуем найти характеристики таких случайных величин, как степень вершины в момент времени t и сумма степеней соседей вершины в момент времени t , а также попробуем сделать предположения о тех характеристиках, которые не получится найти в явном виде. Каждый полученный результат будет проверен с помощью написанной программы для

генерации графа, на основе чего мы будем делать выводы об удачном или неудачном результате.

2 Исследование $M(k_i^n)$

Выражение для математического ожидания степени i -ой вершины хорошо известно и может быть вычислено по следующей формуле:

$$M(k_i(t)) = m\sqrt{\frac{t}{i}}. \quad (1)$$

В данном разделе мы займемся поиском $M(k_i^n(t))$ для произвольного целого n , однако прежде чем приступать к поиску формулы в общем виде, попробуем решить задачу для $n = 2$ и воспользуемся найденным результатом для поиска $D(k_i)$ по формуле:

$$D\xi = M\xi^2 - M^2\xi.$$

2.1 Нахождение $M(k_i^2)$ и $D(k_i)$

Рассматривая величину $k_i^2(i)$ в момент времени $t+1$ и t можно составить разностное уравнение, которое затем можно заменить на следующее приближенное дифференциальное уравнение:

$$\frac{dk_i^2}{dt} = \frac{k_i^2}{t} + \frac{m}{2\sqrt{t \cdot i}}.$$

Решая полученное уравнение, а также пользуясь начальным условием $M(k_i^2(i)) = m^2$, получаем следующий результат:

$$M(k_i^2(t)) = m(m+1)\frac{t}{i} - m\sqrt{\frac{t}{i}}.$$

Этот результат позволяет нам найти дисперсию для $k_i(t)$:

$$D(k_i) = M(k_i^2) - M^2(k_i) = m(m+1)\frac{t}{i} - m\sqrt{\frac{t}{i}} - m^2\frac{t}{i} = m\left(\frac{t}{i} - \sqrt{\frac{t}{i}}\right)$$

Для проверки результата была написана программа на языке Python, генерирующая случайный граф по модели Барабаши-Альберт, а также на каждой итерации считающая нужную нам статистику для выбранной вершины. Сравнение теоретических значений квадрата степени вершины с практически результатами представлено на рис. 1. Сравнение теоретических значений

дисперсии степени вершины с практическими результатами представлено на рис. 2.

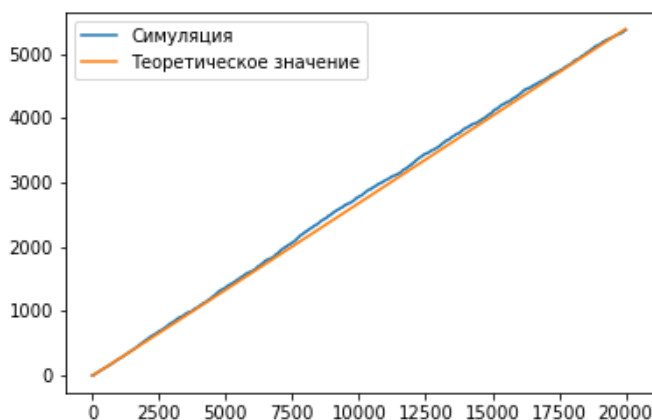


Рисунок 1 – Сравнение теоретического значения $M(k_i^2)$ и практического результата

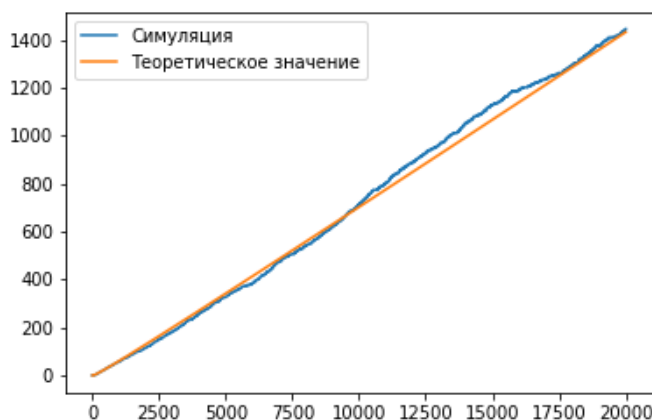


Рисунок 2 – Сравнение теоретического значения для $D(k_i)$ и практического результата

2.2 Вывод формулы для $M(k_i^n)$

Подход, использующий разностное уравнение можно обобщить для произвольной целой степени n величины $k_i(t)$. В этом случае приближенное дифференциальное уравнение имеет следующий вид:

$$\frac{dk_i^n}{dt} = \frac{n \cdot k_i^n}{2t} + \frac{g}{2t},$$

где $g = \sum_{j=0}^{n-2} \binom{n}{j} M(k_i^{j+1})$. Предполагая, что функция g нам известна, то есть известны все математические ожидания меньших степеней, получаем следу-

ющее решение уравнения:

$$M(k_i^n(t)) = \frac{t^{\frac{n}{2}}}{2} \cdot \int t^{-\frac{n}{2}-1} \cdot \sum_{j=0}^{n-2} \binom{n}{j} M(k_i^{j+1}) dt + C \cdot t^{\frac{n}{2}}, \quad (2)$$

где константа C может быть найдена из начального условия $M(k_i^n(i)) = m^n$.

2.3 Нахождение $M(k_i^3)$, $M(k_i^4)$, а также коэффициента асимметрии

Применяя формулу (2) для $n = 3$ получаем следующий результат:

$$M(k_i^3(t)) = m \left(\frac{t}{i}\right)^{\frac{1}{2}} - 3m(m+1)\frac{t}{i} + m(m+1)(m+2) \left(\frac{t}{i}\right)^{\frac{3}{2}},$$

для $n = 4$:

$$M(k_i^4) = m(m+1)(m^2+5m+8) \left(\frac{t}{i}\right)^2 - 6m(m+1)(m+2) \left(\frac{t}{i}\right)^{\frac{3}{2}} + 5m(m+1) \left(\frac{t}{i}\right) - m \left(\frac{t}{i}\right)^{\frac{1}{2}}.$$

Найденное значение $M(k_i^3)$ позволяет нам теперь найти коэффициент асимметрии для случайной величины k_i , который определяется по следующей формуле:

$$\gamma_1 = \frac{\mu_3}{\sigma^3},$$

где σ — стандартное отклонение случайной величины $k_i(t)$, то есть квадратный корень из уже найденной дисперсии, μ_3 — третий центральный момент $k_i(t)$, который может быть вычислен по формуле

$$\mu_3(k_i(t)) = \sum_{j=0}^3 (-1)^j \binom{3}{j} M(k_i^{3-j}(t)) M^j(k_i(t)),$$

Подставляя все известные нам значения, получаем следующую формулу:

$$\gamma_1(k_i) = m^{\frac{1}{2}} \left(\frac{t}{i}\right)^{\frac{1}{4}} \left(\left(\frac{t}{i}\right)^{\frac{1}{2}} - 1 \right)^{\frac{1}{2}} \left(2 \left(\frac{t}{i}\right)^{\frac{1}{2}} - 1 \right) > 0.$$

Результат проверки формулы для $M(k_i^3(t))$ представлен на рис. 3, для $M(k_i^4(t))$ —

на рис. 4.

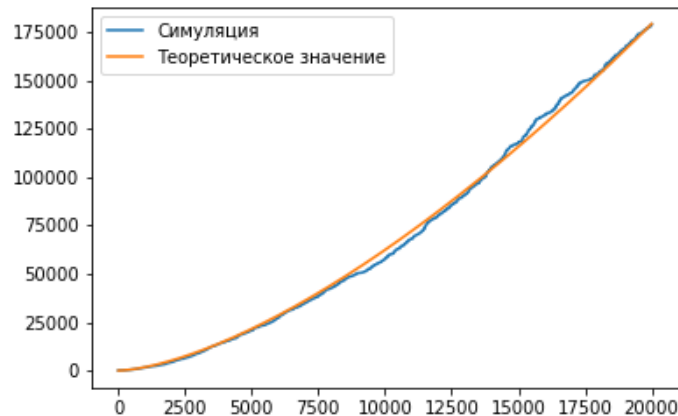


Рисунок 3 – Сравнение теоретического значения для $M(k_i^3)$ и практического результата

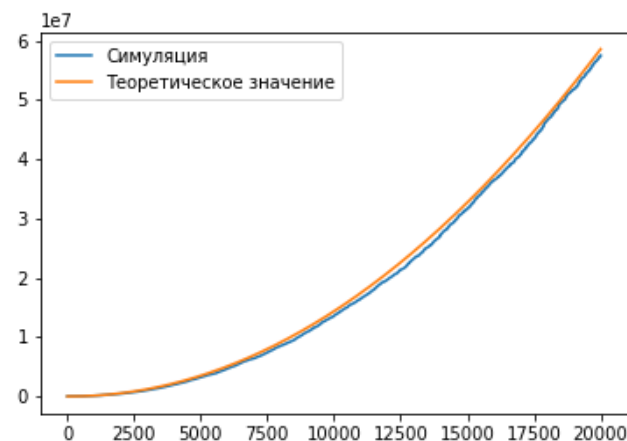


Рисунок 4 – Сравнение теоретического значения для $M(k_i^4)$ и практического результата

Дальнейшее нахождение начальных моментов случайной величины $k_i(t)$ вряд ли будет полезным и интересным занятием ввиду громоздкости получаемых формул уже для $M(k_i^4(t))$, а также в связи с тем, что высшие начальные моменты ничего интересного для исследования степени вершины сказать не смогут, поэтому на этом мы закончим разбор величины $k_i(t)$ и перейдем к другой, но достаточно родственной ей величине.

3 Исследование $M(s_i^2)$

В данном разделе мы немного отойдем от исследования случайной величины k_i , и перейдем к рассмотрению отчасти похожей случайной величины $s_i(t)$ — сумма степеней соседей вершины i в момент времени t . Математическое ожидание этой случайной величины уже было найдено в работе С. П. Сидорова, С. В. Миронова, И. Г. Малинского, Д. Н. Кадомцева «Local degree asymmetry for preferential attachment model» (2021), но в данной работе она будет исследована чуть глубже и будет проанализировано $M(s_i^2)$, что так же, как и в разделе 2.1, позволит нам найти $D(s_i)$. В данной работе мы не будем рассматривать общий вид формулы для $M(s_i^2)$ для любого m , а возьмем лишь частное значение $m = 1$.

3.1 Нахождение точного вида $M(s_i)$

Как уже было сказано ранее, $M(s_i)$ было исследовано в другой работе и был получен следующий результат для $m = 1$:

$$M(s_i) = \frac{1}{2} \left(\frac{t}{i} \right)^{\frac{1}{2}} (\log t + C),$$

где C — некая константа. Попробуем ее найти. Пусть $P(i, j)$ — вероятность того, что вершина i в момент своего появления соединилась с вершиной j , тогда ясно, что $P(i, j) = \frac{k_j(i)}{2t}$. Имеем:

$$M(s_i(i)) = M \left(\sum_{j=1}^{i-1} P(i, j) k_j(i) \right) = M \left(\sum_{j=1}^{i-1} \frac{k_j^2(i)}{2i} \right) = \frac{1}{2i} \sum_{j=1}^{i-1} M(k_j^2(i)).$$

$M(k_j^2(t))$ уже было найдено в 2.1, тогда можем продолжить равенство:

$$\begin{aligned} M(s_i(i)) &= \frac{1}{2i} \sum_{j=1}^{i-1} \left(2 \binom{i}{j} - \sqrt{\frac{i}{j}} \right) = \frac{1}{2i} \left(2i \sum_{j=1}^{i-1} \frac{1}{j} - \sqrt{i} \sum_{j=1}^{i-1} \frac{1}{\sqrt{j}} \right) \approx \\ & \frac{1}{2i} \left(2i \log i - \sqrt{i} \cdot 2\sqrt{i} \right) = \log i - 1, \end{aligned}$$

следовательно,

$$M(s_i(i)) = \frac{1}{2}(\log i + C) = \log i - 1 \implies C = \log i - 2.$$

Окончательно при достаточно больших i имеем:

$$M(s_i) = \frac{1}{2} \left(\frac{t}{i} \right)^{\frac{1}{2}} (\log(t \cdot i) - 2). \quad (3)$$

Полученный результат нам пригодится позднее.

3.2 Составление уравнения для поиска $M(s_i^2)$

Метод поиска формулы не изменился — будем также рассматривать разностное уравнение для исследуемой случайной величины в момент времени t . Введем две вспомогательные вероятности p_1 — вероятность того, что новая вершина присоединилась к вершине i , p_2 — вероятность того, что вершина присоединилась к одному из соседу i . Иными словами:

$$p_1 = \frac{k_i}{2t}$$

$$p_2 = \frac{s_i}{2t}$$

Используя эти вероятности, составляем разностное уравнение для величины $s_i^2(t)$:

$$\begin{aligned} \Delta s_i^2(t+1) &= s_i^2(t+1) - s_i^2(t) = \\ &= p_1(s_i+1)^2 + p_2(s_i+1)^2 + (1-p_1-p_2)s_i^2 - s_i^2 = \\ &= 2s_i(p_1+p_2) + p_1 + p_2 = 2s_i \left(\frac{k_i}{2t} + \frac{s_i}{2t} \right) + \frac{k_i}{2t} + \frac{s_i}{2t} = \\ &= \frac{s_i^2}{t} + \frac{s_i \cdot k_i}{t} + \frac{k_i}{2t} + \frac{s_i}{2t}. \quad (4) \end{aligned}$$

Здесь мы сталкиваемся с проблемой, которой еще не было ранее — мы не можем переходить к математическому ожиданию обеих частей уравнения, так как не все математические ожидания нам известны, в частности нам не известно $M(k_i \cdot s_i)$, а утверждать о независимости s_i и k_i мы не можем, поэтому появляется подзадача — поиск $M(k_i \cdot s_i)$.

3.3 Нахождение $M(k_i \cdot s_i)$

Составляем разностное уравнение для величины $k_i s_i(t)$, которое затем превращаем в следующее приближенное дифференциальное уравнение:

$$\frac{d(k_i s_i)}{dt} = \frac{k_i s_i}{t} + \frac{1}{i}.$$

Полученное уравнение имеет следующее решение:

$$M(k_i s_i) = \frac{t \log t}{i} + Ct,$$

где C — константа интегрирования, которую нам предстоит найти.

Рассмотрим $M(k_i s_i)$ в момент времени i : так как $M(k_i(i))$ — по сути, константа, равная 1, то получаем, что $M(k_i s_i) = M(s_i(i)) = \log i - 1$, следовательно,

$$\log i + Ci = \log i - 1 \implies C = -\frac{1}{i}.$$

Окончательно имеем:

$$M(k_i s_i) = \frac{t \log t}{i} - \frac{t}{i} = \frac{t}{i}(\log t - 1).$$

3.4 Нахождение $M(s_i^2)$ и $D(s_i)$

Теперь можем вернуться к уравнению (4) и перейти к математическому ожиданию обеих частей, так как на этот раз нам все известно, получаем:

$$\frac{ds_i^2}{dt} = \frac{s_i^2}{t} + \frac{\log t - 1}{i} + \frac{\log(t \cdot i)}{4\sqrt{t \cdot i}}.$$

Данное уравнение имеет следующее решение:

$$M(s_i^2) = \frac{t}{i} \log t \left(\frac{\log t}{2} - 1 \right) - \left(\frac{t}{i} \right)^{\frac{1}{2}} \left(1 + \frac{\log(t \cdot i)}{2} \right) + Ct,$$

где C — константа интегрирования, которая может быть найдена, используя тот же подход, с помощью которого была найдена константа для $M(s_i(t))$:

$$C = \frac{2 - \log i + 3\sqrt{i} \cdot B - \frac{\log^2 i}{2}}{i},$$

где $B = \sum_{j=1}^{\infty} \frac{1}{j^{\frac{3}{2}}} \approx 2.6124$.

Данная формула была проверена с помощью серии экспериментов с параметрами $T = 300$, $i = 50$. Результат представлен на рис. 5.

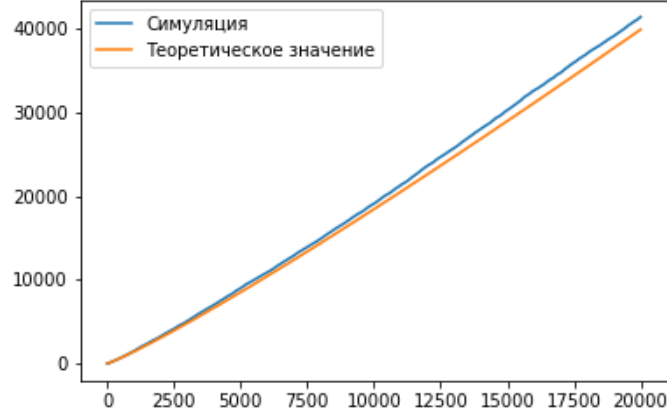


Рисунок 5 – Сравнение теоретического значения $M(s_i^2)$ и практического результата

Полученный результат позволяет нам найти $D(s_i)$:

$$D(s_i) = M(s_i^2) - M^2(s_i) = \left(\frac{t}{i}\right) \left(\frac{\log^2 t}{2} - \log t - \frac{1}{4}(\log(t \cdot i) - 2)^2\right) - \left(\frac{t}{i}\right)^{\frac{1}{2}} \left(1 + \frac{\log(t \cdot i)}{2}\right) + Ct,$$

где

$$C = \frac{2 - \log i + 3\sqrt{i} \cdot B - \frac{\log^2 i}{2}}{i}.$$

3.5 Проблемы с нахождением $M(s_i)$ для произвольного m

Возникает вопрос, можем ли мы с помощью тех же самых уравнений попробовать найти $M(s_i)$ для любого m ? Эта задача оказывается на порядок труднее, чем та, что была решена для $m = 1$, поскольку теперь, составляя разностное уравнение, нам придется учитывать изменение $s_i(t)$ на всевозможные значения, а не только на 0 и 1, причем вместе с этим появятся сложные вероятности, связанные с событиями увеличения s_i на каждую величину.

Однако никто не мешает провести эксперимент и попробовать хотя бы эмпирически оценить закон, которой будет подчиняться $M(s_i)$. Глядя на урав-

нение 3, можно сделать предположение, что

$$\forall m \mathbb{M}(s_i(t)) \sim \frac{1}{2} C(m) \left(\frac{t}{i}\right)^{\frac{1}{2}} \cdot \log t \text{ при } t \rightarrow \infty,$$

где C — некая константа, зависящая только от m . Однако была проведена серия экспериментов, которая, к сожалению, данное предположение опровергла.

ЗАКЛЮЧЕНИЕ

Настоящая работа начата в рамках анализа, проведенного в работе С. П. Сидорова, С. В. Миронова, И. Г. Малинского, Д. Н. Кадомцева «Local degree asymmetry for preferential attachment model» (2021), и явилась его логическим продолжением.

В данной работе были исследованы такие случайные величины, связанные с моделью Барабаши-Альберт, как степень выбранной вершины и сумма степеней соседей выбранной вершины, была написана программа для генерации графа и подсчета исследуемой нами величины.

В ходе написания работы были решены следующие задачи:

- было найдено приближенное значение для начального момента случайной величины $k_i(t)$ произвольного порядка.
- опираясь на найденные начальные моменты, были найдены такие характеристики случайной величины $k_i(t)$ как дисперсия и коэффициент асимметрии;
- взяв частный случай $m = 1$, для случайной величины $s_i(t)$ удалось найти приближенный вид ее начальных моментов первого и второго порядка, используя которые получилось найти ее дисперсию.
- было предположено поведение математического ожидания $s_i(t)$ при любом m ;
- все полученные формулы удалось подтвердить эмпирически с помощью ранее написанной программы для генерации графа модели Барабаши-Альберт.

В качестве открытых вопросов первую очередь хочется упомянуть проблему нахождения $M(s_i(t))$ для произвольного числа ребер m . Возможно к этой задаче нужен подход, отличный от того, с которым мы подходили ко всем задачам в этой работе, связанный с рассмотрением разностного уравнения. Решение этой задачи позволит продвинуться дальше и поговорить о таком вопросе, как парадокс дружбы в графе модели Барабаши-Альберт. Парадокс дружбы на простом языке говорит о том, что в среднем друзья одного человека популярнее него самого, если под популярностью подразумевать количество друзей. На языке используемых нами величин этот парадокс можно записать

как:

$$M\left(\frac{s_i(t)}{k_i(t)}\right) > M(k_i(t)).$$

Этот вопрос уже был рассмотрен в работе «Local degree asymmetry for preferential attachment model» (2021) и даже было доказано, что при $m = 1$ парадокс в графе данной модели наблюдается, однако будет ли он наблюдаться при любых m ? А если мы позволим m быть дискретной случайной величиной? Кстати последнее усложнение также стоит отдельного изучения, поскольку достаточно сильно приближает модель из реального мира, когда, например, количество ссылок в статье или на веб-странице — неконстантная величина.