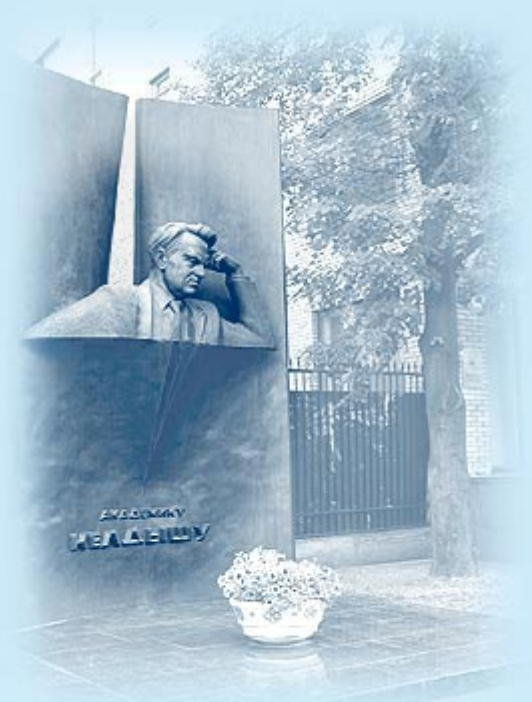




ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 88 за 2022 г.



ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

А.А. Кислицын, М.В. Гогуев

**Вычислительная экспансия в
графы ближайших соседей:
статистики и размерности
пространства**

Статья доступна по лицензии
[Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/)



Рекомендуемая форма библиографической ссылки: Кислицын А.А., Гогуев М.В. Вычислительная экспансия в графы ближайших соседей: статистики и размерности пространства // Препринты ИПМ им. М.В.Келдыша. 2022. № 88. 32 с. <https://doi.org/10.20948/prepr-2022-88>
<https://library.keldysh.ru/preprint.asp?id=2022-88>

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

А.А. Кислицын, М.В. Гогуев

**Вычислительная экспансия в графы
ближайших соседей: статистики
и размерности пространства**

Москва — 2022

Кислицын А.А., Гогуев М.В.

Вычислительная экспансия в графы ближайших соседей: статистики и размерности пространства

Исследуются распределения графов первых ближайших соседей по числу несвязных фрагментов, фрагментов по числу вершин и вершин по степеням входящих ребер в зависимости от числа вершин графа. Рассмотрены две ситуации: когда матрица случайных расстояний задается непосредственно и когда задаются случайные координаты точек в евклидовом пространстве данной размерности. В ходе вычислительного эксперимента показано, что для размерностей выше пятой распределения вершин по степеням в первом и втором вариантах могут использоваться на близких уровнях значимости.

Ключевые слова: граф ближайших соседей, распределение вершин по степеням, распределение расстояний между точками

Kislitsyn A.A., Goguev M.V.

Computational expansion into nearest neighbor graphs: statistics and dimensions of space

The distributions of graphs of the first nearest neighbors by the number of disconnected fragments, fragments by the number of vertices, and vertices by the degrees of incoming edges, depending on the number of vertices of the graph, are investigated. Two situations are considered: when the matrix of random distances is given directly, and when random coordinates of points in Euclidean space of a given dimension are given. In the course of a computational experiment, it is shown that with an increase in the dimension of the space, the statistics of the first and second variants converge. For dimensions above the fifth, the degree distributions of the vertices could be used approximately at the same significance level.

Keywords: graph of nearest neighbors, distribution of vertices by degrees, distribution of distances between points

Работа выполнена при поддержке гранта РФФИ № 19-71-30004

Содержание

1. Введение	3
2. Распределение расстояний между точками	5
3. Алгоритм генерации матрицы попарных расстояний	8
4. Визуализация графов ближайших соседей	11
5. Результаты вычислительных экспериментов	13
6. Заключение	31
Литература	31

1. Введение

Настоящая работа продолжает исследования по статистике графов, начатые в [1-2]. Цель работы состоит в формализации выборочного статистического метода применительно к графам ближайших соседей.

Метод k -х ближайших соседей (k nearest neighbors, k -NN – см., напр., [3]) представляет собой один из широко используемых методов кластеризации данных в метрическом пространстве. В большинстве случаев целью анализа является установление структурных связей между объектами, которым были поставлены в соответствие наборы чисел и введена мера близости. То есть в задаче кластеризации данных сами эти числа, как и расстояния между ними, интереса не представляют, важным является результат группировки чисел, позволяющий выявить свойства взаимного расположения объектов. Кластеризация – это эффективный способ сокращения описания, когда вместо большого количества исходных объектов оказывается возможным рассмотреть относительно небольшое число классов и сформировать типичного представителя класса.

Мотивация предпринятого исследования обусловлена трудностями, возникающими при попытке систематизации графов ближайших соседей в контексте статистического анализа свойств исходного набора объектов.

Первая проблема состоит в том, что понятие статистической выборки относится к группе объектов, а не к представляющему эту группу графу. Если бы таких групп было достаточно для построения их выборочной функции распределения по тем или иным параметрам, то можно было бы сопоставить, допустим, квантилям такого распределения структуру графов k -NN. Однако группа объектов, как правило, в исследовании всего одна. Например, множество полевых цветов на участке. А другая группа объектов может относиться совсем к другой категории – допустим, к покупателям в магазине. Результаты отдельных исследований таких групп не могут быть объединены, так как они относятся к разным распределениям. Вопрос же, на который требуется ответить, состоит в том, насколько типичной для анализируемого множества (цветов, людей и т.п.) является наблюдаемая структура графа. Для этого надо от одного случайного графа k -NN перейти к набору таких графов.

Теория случайных графов развита к настоящему времени весьма детально. Начиная с классической работы [4] свойства графов, вершины которых соединяются между собой по заданному вероятностному закону, изучались в контексте решения многих прикладных задач (см., напр., монографии [5-7]). Однако в том, что касается математической статистики графов k -NN, конструктивных моделей и результатов существенно меньше. Это связано с тем, что «выборка графов» сама по себе не имеет места, ведь граф просто визуализирует некоторое свойство изучаемой группы объектов. Так что понятие выборки адресуется именно группе объектов, которые должны быть однородными по изучаемому свойству в статистическом смысле. Но поскольку

изучаемое свойство выявляется как раз в результате анализа структуры графа, заранее сформировать группы объектов с нужным свойством невозможно. Тогда, рассмотрев распределение параметров, которые в той или иной модели системы считаются ключевыми для конкретной выборки, можно размножить выборки с найденным эмпирическим распределением параметров, после чего изучить статистику графов, отвечающих (по предположению модели) различным состояниям изучаемой системы.

Замечательным свойством графов k -NN является то, что распределение их структур не зависит от распределения расстояний между объектами [1]. Это позволяет организовать вычислительный эксперимент по построению статистик графов k -NN. Но тут возникает вторая проблема: что именно генерировать? Можно задать равномерное распределение координат точек в некотором n -мерном евклидовом пространстве, но тогда надо перебирать всевозможные значения размерности, поскольку структура графа от нее, очевидно, зависит. Например, нельзя ожидать третьей степени вершины по входящему ребру для графа ближайших соседей на прямой и шестой степени на плоскости. В то же время размерность пространства, в котором размещено изучаемое множество объектов, как правило, заранее не известна.

С другой стороны, можно сгенерировать матрицу случайных расстояний между точками и анализировать графы k -NN непосредственно по этим данным. Такой способ наиболее удобен, но возникает третья проблема. Для любых трех точек сгенерированные расстояния между ними должны удовлетворять неравенству треугольника, что требует специального алгоритма, который становится затратным для большого набора точек. В настоящей работе будет представлен такой алгоритм и описаны полученные с его помощью результаты статистического анализа распределений основных статистик графов первых ближайших соседей. Именно, изучаются статистики распределения графов по числу несвязных фрагментов, фрагментов по числу вершин, вершин по степеням входящих ребер. Кроме того, будет проведено сравнение получающихся статистик с результатами генерации случайного графа ближайших соседей непосредственно через координаты точек, заданных в евклидовом пространстве определенной размерности. Поскольку этот второй способ генерации случайного графа вычислительно более затратный, причем эти затраты растут с увеличением размерности, то возможность построения статистик на достаточном уровне значимости без учета размерности пространства представляется практически важной.

Мы будем предполагать, что все попарные расстояния между точками различны, то есть для каждой точки существует единственный ближайший сосед. Тогда графы 1-NN имеют следующие основные свойства:

- с каждой вершиной связано ровно одно выходящее ребро, поэтому изолированных точек в графе 1-NN нет;
- число вершин графа равно числу ребер;
- сумма степеней вершин равна числу вершин;
- каждый связный граф 1-NN содержит ровно один цикл длины два.

В работах [1, 2] было показано, что для при моделировании матрицы расстояний между точками безотносительно размерности пространства число несвязных фрагментов графа первых ближайших соседей имеет распределение, асимптотически имеющее вид полукругового распределения Вигнера для $[N/2]$, где N есть число вершин графа. В пространствах определенной размерности также следует провести соответствующие численные эксперименты.

2. Распределение расстояний между точками

Так как статистика графов 1-NN не зависит от распределения расстояний между точками, можно задать любое распределение. Смысл этого утверждения следующий. Каждая конкретная матрица расстояний порождает определенный граф 1-NN. С этой матрицей связано определенное эмпирическое распределение расстояний. Возьмем вторую матрицу расстояний, третью и т.д. Если взять таких матриц достаточно много, например миллион, то статистика структур получаемых графов не будет зависеть от того, какие распределения расстояний получались для этих матриц. Можно даже считать, что каждое распределение расстояний отлично от других. Удобно считать, что используемая на практике дискретная модель распределения расстояний аппроксимирует некоторую непрерывную функцию распределения, что позволяет доказать ряд утверждений.

Поскольку любое непрерывное распределение генерируется с помощью равномерного распределения, достаточно рассмотреть статистику графов для матрицы случайных попарных расстояний, равномерно распределенных на отрезке $[0; 1]$. Однако именно такое распределение расстояний невозможно реализовать на практике. Ниже приведено доказательство этого утверждения.

Предложение 1. Каково бы ни было распределение координат точек с интегрируемым квадратом плотности функции распределения на ограниченном замкнутом множестве в евклидовом пространстве конечной размерности, распределение расстояний между точками не может быть равномерным.

Доказательство. Рассмотрим сначала распределение точек в R^1 . Без ограничения общности ограниченное замкнутое множество считаем отрезком $[0; 1]$. Пусть на нем задано распределение координат x точек с плотностью

$$f(x) \geq 0, \int_0^1 f(x) dx = 1. \text{ Рассмотрим на том же множестве вторую случайную}$$

величину с распределением $g(y)$, так что $g(y) \geq 0, \int_0^1 g(y) dy = 1$. Плотность

распределения разности $z = x - y$ этих случайных величин дается сверткой

$$h(z) = \int_0^1 f(z + y)g(y)dy \quad (2.1)$$

с областью определения $z \in [-1; 1]$. При $z \geq 0$ функция $f(z + y)$ равна нулю, если $z + y > 1$, так что y меняется от 0 до $1 - z$. При $z < 0$ функция $f(z + y)$ равна нулю, если $z + y < 0$, так что y меняется от $-z$ до 1. В результате свертка (2.1) представляется в виде суммы двух интегралов для положительных и отрицательных значений разности $x - y$:

$$\begin{aligned} h(z) &= \int_0^{1-z} f(z+y)g(y)dy + \int_{-z}^1 f(z+y)g(y)dy = \\ &= \int_0^{1-z} f(z+y)g(y)dy + \int_0^{1+z} f(y)g(y-z)dy \equiv \varphi^+(z) + \varphi^-(z). \end{aligned} \quad (2.2)$$

Поэтому распределение $\rho(r)$ модуля разности $r = |x - y|$ дается формулой

$$\rho(r) = \varphi^+(z) + \varphi^-(-z), \quad r = |z|. \quad (2.3)$$

Тогда из (2.2) и (2.3) получаем

$$\rho(r) = \int_0^{1-r} f(y+r)g(y)dy + \int_0^{1-r} f(y)g(y+r)dy, \quad r \in [0; 1]. \quad (2.4)$$

Отсюда, в частности, следует, что модуль разности двух независимых одинаково распределенных случайных величин (когда $f = g$) дается формулой

$$\rho(r) = 2 \int_0^{1-r} f(x)f(x+r)dx, \quad r \in [0; 1]. \quad (2.5)$$

Плотность распределения модуля разности неотрицательна и также нормирована на единицу на отрезке $[0; 1]$. При заданной функции $\rho(r)$ уравнение (2.5) представляет собой интегральное уравнение с квадратичной нелинейностью. Отдельные примеры решений таких уравнений содержатся в справочнике [8, с. 302-304], однако общей теории пока не развито. В классе рассматриваемых функций $f(x)$ оказывается, что отнюдь не любая функция $\rho(r)$ возможна. В частности, равномерное распределение расстояний между точками в R^1 не может быть реализовано как модуль разности одинаково распределенных случайных величин. Действительно, в этом случае должно быть $\rho(r) = 1 \quad \forall r \in [0; 1]$. Но при $r \rightarrow 1$ интеграл в правой части (2.5) стремится к нулю, поэтому распределение расстояний (2.5) заведомо не является равномерным.

Рассмотрим теперь расстояние между точками в R^2 в квадрате $[0; 1]^2$. Пусть по первой координате плотность распределения дается функцией f , а по второй – функцией g . Согласно (2.1), разности $u = x - x'$ и $v = y - y'$ имеют плотности распределения

$$h_1(u) = \int_0^1 f(u+x)f(x)dx, \quad h_2(v) = \int_0^1 g(u+y)g(y)dy. \quad (2.6)$$

Случайные величины $z = u^2$ и $s = v^2$ имеют плотности распределения $\psi_1(z) = h_1(\sqrt{z})/(2\sqrt{z})$ и $\psi_2(s) = h_2(\sqrt{s})/(2\sqrt{s})$ соответственно, т.е.

$$\psi_1(z) = \frac{1}{2\sqrt{z}} \int_0^1 f(\sqrt{z} + x) f(x) dx, \quad \psi_2(s) = \frac{1}{2\sqrt{s}} \int_0^1 g(\sqrt{s} + x) g(x) dx. \quad (2.7)$$

Тогда случайная величина $w = z + s \equiv u^2 + v^2$ имеет плотность распределения

$$\begin{aligned} \chi(w) &= \int_0^w \psi_1(w-s) \psi_2(s) ds = \\ &= \frac{1}{4} \int_0^w \frac{ds}{\sqrt{(w-s)s}} \int_0^1 f(\sqrt{w-s} + x) f(x) dx \cdot \int_0^1 g(\sqrt{s} + y) g(y) dy. \end{aligned} \quad (2.8)$$

Таким образом, из (2.8) следует, что случайная величина $r = \sqrt{w} = \sqrt{u^2 + v^2}$ имеет плотность распределения

$$\rho(r) = 2r\chi(r^2) = \frac{r}{2} \int_0^{r^2} \frac{ds}{\sqrt{(r^2-s)s}} \int_0^1 f(\sqrt{r^2-s} + x) f(x) dx \cdot \int_0^1 g(\sqrt{s} + y) g(y) dy. \quad (2.9)$$

Сделаем в (2.9) замену переменных $t = s/r^2$. Тогда получаем:

$$\rho(r) = \frac{r}{2} \int_0^1 \frac{dt}{\sqrt{t(1-t)}} \int_0^1 f(r\sqrt{1-t} + x) f(x) dx \cdot \int_0^1 g(r\sqrt{t} + y) g(y) dy. \quad (2.10)$$

Поскольку плотности f и g квадратично интегрируемы, то интегралы

$$\int_0^1 f(r\sqrt{1-t} + x) f(x) dx \quad \text{и} \quad \int_0^1 g(r\sqrt{t} + y) g(y) dy$$

отличаются соответственно от $\int_0^1 f^2(x) dx$ и $\int_0^1 g^2(y) dy$ на величины порядка $O(r)$. Тогда получаем

$$\rho(r) = \frac{r}{2} \int_0^1 \frac{dt}{\sqrt{t(1-t)}} \cdot \left(\int_0^1 f^2(x) dx + O(r) \right) \cdot \left(\int_0^1 g^2(y) dy + O(r) \right).$$

Поскольку интеграл $\int_0^1 \frac{dt}{\sqrt{t(1-t)}}$ вычисляется явно (он равен π), то для интеграла

(2.10) получаем оценку

$$\rho(r) = \frac{\pi r}{2} \int_0^1 f^2(x) dx \cdot \int_0^1 g^2(y) dy + o(r). \quad (2.11)$$

В результате при $r \rightarrow 0$ из (2.11) получаем $\rho(r) \rightarrow 0$. Поэтому плотность $\rho(r)$ не может быть плотностью равномерно распределенной величины.

Отсюда по индукции заключаем, что на множестве $[0; 1]^n$ невозможно получить равномерно распределенное расстояние между точками, так как распределение случайной величины $r_n = \sqrt{r_{n-1}^2 + v^2}$ строится таким же образом, как и распределение величины r_{n-1} , по формулам, аналогичным (2.8)-(2.11).

Предложение 1 доказано.

Тем не менее, поскольку упорядоченность расстояний не зависит от функции распределения этих расстояний, удобно использовать именно равномерное распределение для генерации матрицы случайных расстояний.

3. Алгоритм генерации матрицы попарных расстояний

Перейдем теперь к описанию алгоритма генерации матрицы r_{ij} случайных расстояний между точками i, j .

Пусть требуется построить матрицу расстояний между N точками в евклидовом пространстве. Поскольку эта матрица симметрична, достаточно сгенерировать $N(N-1)/2$ случайных величин на отрезке $[0;1]$. Первые $N-1$ точек будут отвечать расстояниям r_{1j} , $j=2,3,\dots,N$, следующие $N-2$ точек – расстояниям r_{2j} , $j=3,4,\dots,N$ и т.д. Однако для анализа, например, циклических подграфов в графе k -NN следует убедиться, что расстояния удовлетворяют неравенству треугольника:

$$\forall i, j, k : r_{ij} + r_{jk} \geq r_{ik}. \quad (3.1)$$

В противном случае некоторые циклы могут оказаться невозможны.

Как известно (см., напр., [9]), вероятность того, что три случайных числа отвечают длинам некоторого треугольника, равна $1/4$. Поскольку выбрать три значения из $S = N(N-1)/2$ можно C_S^3 способами, то вероятность того, что сгенерированная матрица есть матрица расстояний, равна $4^{-C_S^3}$, что при больших N имеет порядок $4^{-N^6/48}$. Следовательно, если просто генерировать произвольные матрицы из равномерно распределенных чисел, а затем отобрать те, которые обладают требуемым свойством, то даже для такого малого числа, как $N=100$, только одна на 4 в степени 20 миллиардов будет подходящей. Это, разумеется, совершенно неприемлемо с вычислительной точки зрения. Тогда надо пойти другим путем и сразу генерировать матрицу с нужным свойством. Опишем соответствующий алгоритм построения.

На первом шаге генерируются $N-1$ чисел, равномерно распределенных на отрезке $[0;1]$. Поскольку нумерация точек не имеет значения, то занумеруем их в порядке возрастания полученных расстояний, положив наименьшее из них за расстояние r_{12} . Следующее расстояние будет отвечать паре r_{13} и т.д.

На втором шаге строим $N-2$ расстояний для второй точки. Расстояние r_{23} должно удовлетворять условиям $r_{12} + r_{23} \geq r_{13}$ и $r_{13} + r_{12} \geq r_{23}$. Поэтому r_{23} есть некоторое число между $r_{13} - r_{12}$ и $r_{13} + r_{12}$. При этом условие $r_{13} + r_{23} \geq r_{12}$ выполняется автоматически, так как $r_{23} \geq r_{13} - r_{12}$, а тогда $r_{13} + r_{23} \geq 2r_{13} - r_{12}$, что больше, чем r_{12} в силу того, что на первом шаге $r_{13} > r_{12}$. Аналогично расстояние r_{24} должно удовлетворять условиям $r_{14} - r_{12} \leq r_{24} \leq r_{14} + r_{12}$ и т.д.

Следовательно, надо построить промежутки $\Delta_{1,2}^{1,j} = [r_{1j} - r_{12}; r_{1j} + r_{12}]$, после чего сгенерировать по одной случайной точке r_{2j} на каждом из них.

На третьем шаге строим $N - 3$ расстояний для третьей точки. Первое из них r_{34} должно принадлежать пересечению отрезков $\Delta_{1,3}^{1,4}$ и $\Delta_{2,3}^{2,4}$. Это пересечение заведомо не пусто. Действительно, в противном случае либо $r_{14} + r_{13} < |r_{24} - r_{23}|$, либо $r_{24} + r_{23} < |r_{14} - r_{13}|$. Рассмотрим, например, первый вариант. Пусть $r_{24} > r_{23}$. Тогда должно быть $r_{14} + r_{13} + r_{23} < r_{24}$. Но так как $r_{24} + r_{14} \geq r_{21}$ из второго шага, то получается, что $2r_{14} + r_{13} + r_{23} < r_{21}$. Но также из второго шага следует, что $r_{13} + r_{23} \geq r_{21}$, что противоречит предыдущему неравенству. Пусть теперь $r_{24} < r_{23}$. Тогда должно быть $r_{14} + r_{13} + r_{24} < r_{23}$. Но в силу построения на втором шаге $r_{14} + r_{24} \geq r_{21}$, и потому левая часть получившегося неравенства больше, чем $r_{13} + r_{21}$, что, в свою очередь, больше, чем r_{23} , тогда как по предполагаемому условию левая часть меньше, чем r_{23} . Оставшийся вариант рассматривается аналогично. Таким образом, надо построить пересечение известных промежутков $\delta_3^j = \Delta_{1,3}^{1,j} \cap \Delta_{2,3}^{2,j}$ и выбрать на них по одной произвольной точке. Это и будут расстояния r_{3j} .

На следующем шаге область значений расстояний r_{4j} , $j > 4$, должна будет принадлежать пересечению трех условий $\Delta_{1,4}^{1,j} \cap \Delta_{2,4}^{2,j} \cap \Delta_{3,4}^{3,j}$. Покажем, что с увеличением числа пересечений соответствующее множество не пусто.

Предложение 2. Пусть r_{ij} – случайные расстояния между точками i и j , причем нумерация N точек задана по возрастанию первых $N-1$ расстояний $r_{12}, \dots, r_{1N}$. Определим отрезки $\Delta_{k,i}^{k,j} = [r_{kj} - r_{ki}; r_{kj} + r_{ki}]$. Тогда $\forall i > 2$ пересечение отрезков $\Delta_{1,i}^{1,j} \cap \Delta_{2,i}^{2,j} \dots \cap \Delta_{i-1,i}^{i-1,j}$ не пусто.

Доказательство. Для $i = 2$ и $i = 3$ утверждение выполнено, как показано выше. Пусть предположение 2 выполнено для всех расстояний $r_{ij}, j > i$. Рассмотрим расстояние $r_{i+1, i+2}$. Для него должны выполняться условия:

$$\begin{aligned} & \left| r_{1,i+2} - r_{1,i+1} \right| \leq r_{i+1,i+2} \leq r_{1,i+2} + r_{1,i+1}; \\ & \left| r_{2,i+2} - r_{2,i+1} \right| \leq r_{i+1,i+2} \leq r_{2,i+2} + r_{2,i+1}; \\ & \dots\dots\dots \\ & \left| r_{i,i+2} - r_{i,i+1} \right| \leq r_{i+1,i+2} \leq r_{i,i+2} + r_{i,i+1}. \end{aligned} \tag{3.2}$$

По предположению индукции пересечение $\Delta_{1,i}^{1,j} \cap \Delta_{2,i}^{2,j} \dots \cap \Delta_{i-1,i}^{i-1,j}$ не пусто. Предположим, что пересечение последнего отрезка в (3.2) с оставшимися отрезками пусто. При этом первые $i-1$ отрезков имеют непустое пересечение по предыдущему построению. Предположим, что левый конец последнего отрезка больше правого конца предпоследнего отрезка, т.е.

$|r_{i,i+2} - r_{i,i+1}| > r_{i-1,i+1} + r_{i-1,i+2}$. Допустим, что $r_{i,i+1} > r_{i,i+2}$. Тогда для выполнения нашей противоположной гипотезы должно выполняться неравенство $r_{i,i+1} > r_{i-1,i+1} + r_{i-1,i+2} + r_{i,i+2}$. Поскольку на предыдущем шаге было получено условие $r_{i-1,i+2} + r_{i,i+2} \geq r_{i-1,i}$, то проверяемое контр-условие принимает вид $r_{i,i+1} > r_{i-1,i+1} + r_{i-1,i}$. Но на том же предыдущем шаге было выяснено, что $r_{i-1,i+1} + r_{i-1,i} > r_{i,i+1}$, что противоречит контр-условию. Значит, предположение о пустом пересечении неверно. Остальные варианты – когда левый конец последнего отрезка больше правого конца остальных отрезков системы (3.2) или когда правый конец последнего отрезка меньше левого конца любого из отрезков системы (3.2), рассматриваются аналогично. **Предложение 2 доказано.**

Отсюда следует, что генерация матрицы случайных расстояний может быть построена следующим образом.

Шаг 1. Генерируются $N(N-1)/2$ чисел, равномерно распределенных на отрезке $[0;1]$. Обозначим эти числа r'_{ij} . Обозначим также r_{ij} числа, которые отвечают случайным расстояниям в соответствующей матрице.

Шаг 2. Первые $N-1$ чисел упорядочиваются по возрастанию, после чего происходит нумерация точек в указанном порядке по отношению к произвольно взятой первой точке. Полагаем для этих чисел $r_{1j} = \hat{T}r'_{1j}$, где $\hat{T}$ есть оператор упорядочения по возрастанию. Так как нас интересует структура графа с точностью до изоморфизма, то от нумерации точек статистика степеней вершин, длин циклов и т.п. не зависит.

Шаг 3. Определяются промежутки $\Delta_{1,2}^{1,j} = [r_{1j} - r_{12}; r_{1j} + r_{12}] \equiv [\alpha_2^j; \beta_2^j]$. Для следующих $N-2$ точек $r'_{2j}, j > 2$ делается преобразование, перемещающее точку r'_{2j} в отрезок $[\alpha_2^j; \beta_2^j]$:

$$r_{2j} = \alpha_2^j(1 - r'_{2j}) + \beta_2^j r'_{2j} = \alpha_2^j + (\beta_2^j - \alpha_2^j)r'_{2j}. \quad (3.3)$$

Шаг 4. Для найденных точек r_{2j} аналогично шагу 2 определяются промежутки $\Delta_{2,3}^{2,j} = [\alpha_{2,3}^j; \beta_{2,3}^j]$, после чего находятся концы отрезков $[a_3^j; b_3^j]$, $j > 3$ пересечения условий $\delta_3^j = \Delta_{1,3}^{1,j} \cap \Delta_{2,3}^{2,j}$. Затем определяются точки r_{3j} по формуле, аналогичной (3.3): $r_{3j} = a_3^j(1 - r'_{3j}) + b_3^j r'_{3j}$, и т.д.

В результате получается набор из $N(N-1)/2$ расстояний, которые удовлетворяют неравенству треугольника. Однако, как следует из данного алгоритма, они уже не являются равномерно распределенными. Хотя, как уже говорилось выше, вид распределения не принципиален, его можно привести к равномерному на отрезке $[0;1]$. Пусть $f(r)$ есть плотность этого

распределения, а $F(r) = \int_0^r f(x)dx$ есть его функция распределения. Тогда в силу

монотонности отображения $y_k = F(r_k)$ упорядоченность чисел y_k на $[0; 1]$ такая же, как и чисел r_k , но по построению числа y_k имеют равномерное распределение на $[0; 1]$.

4. Визуализация графов ближайших соседей

Существует большое количество алгоритмов, реализующих визуализацию метода ближайших соседей: так называемое дерево вложенных шаров [10], k-d-дерево [11], алгоритм ближайших контуров [12], дерево основных осей [13] и другие (см., напр., обзор алгоритмов метода k-NN в [14]).

Для проведения расчетов было сгенерировано 5 файлов, каждый из которых представляет собой 1 млн векторов длиной 1000 случайных значений с плавающей точкой, равномерно распределенных на интервале $[0, 1)$ с точностью до 17 знака после запятой. Для удобства организации расчетов каждый файл был поделен на 10 одинаковых частей. Каждая часть содержит соответственно 100 000 векторов по 1000 элементов. Данные сохраняются в сжатом .prz файле. Размер одного такого файла ~ 740 МБ.

Для пространств 1D–5D мы имеем группы файлов, которые характеризуют по одной размерности в этих пространствах. Далее для каждого вектора в группе файлов определенной размерности необходимо построить матрицу расстояний между элементами в рассматриваемом пространстве. Расстояние вычисляется как евклидово расстояние между точками.

В итоге мы построим для каждого вектора симметричную матрицу расстояний 1000×1000 , которую преобразуем в матрицу первого ближайшего соседа.

Для создания графа из получившейся матрицы воспользуемся функционалом библиотеки NetworkX, предназначенной для создания, манипуляции и изучения структуры, динамики и функционирования сложных сетевых структур [<https://networkx.org/>]. За счет того, что библиотека реализована на низкоуровневой структуре данных языка Python, обеспечивается эффективное расходование аппаратных и программных ресурсов. Это позволяет оптимально рассчитать интересующие нас статистики по подграфам в каждой реализации графа и сохранить их в сжатом виде.

Для ускорения процесса расчета была реализована параллельная версия алгоритма. Так как 10 частей, на которые разбита исходная размерность, независимы друг от друга, то можно распараллелить расчет, раздав эти части на обработку разным потокам.

Параллельная реализация считалась на 10 ядрах процессора Intel Core i9 10920X. Данные по времени расчета различных размерностей представлены в Таблице 1.

Таблица 1. Зависимость времени расчета графа от размерности

Размерность	ч : м : с
D1	33:38:56

D2	48:38:27
D3	65:43:56
D4	79:20:44
D5	96:19:03

Как уже было описано выше, результаты расчетов были получены для размерностей 1D-5D пространства. Для этих вариантов результирующие статистики были сохранены в 100 файлах типа pickle. Этот модуль сохранения данных был выбран благодаря тому, что он позволяет оптимально и однозначно преобразовывать python-объекты в бинарные структуры и обратно. Каждый из таких файлов содержит статистики по 10 000 графам, включающие количество подграфов, распределение вершин по степеням и количество таких вершин. Данные хранятся максимально компактно в python-объекте, однако даже при таком хранении данных размер одного результирующего файла ~ 181 МБ. Итого, для хранения статистических данных для 1 млн реализаций графа на 1000 вершин необходимо порядка 18 ГБ.

Перед визуализацией данных происходит их подготовка. Суммируются статистики по степеням вершин и суммируются вершины по компонентам подграфа. Далее строится их распределение и сохраняется в матрицы A [Мах степень вершины в графе, Мах суммарное кол-во вершин определенной степени в реализации графа] и Z [Мах кол-во вершин в компоненте подграфа, Мах кол-во подграфов в графе].

Визуализация результатов расчетов проводилась с использованием программных средств библиотеки plotly [<https://plotly.com/python>]. Пример визуализации графа случайных первых ближайших соседей для пространства 3D приведен на рис. 1.

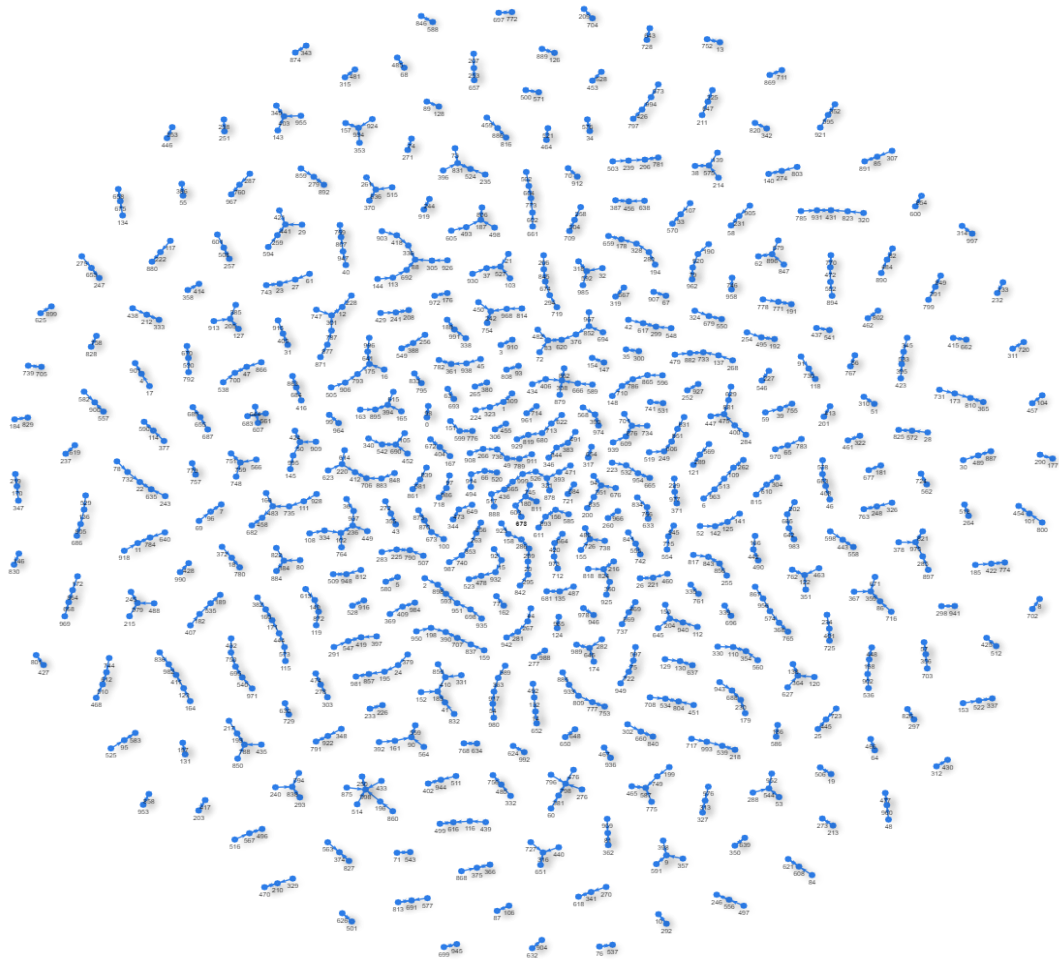


Рис. 1 – Пример направленного графа первого ближайшего соседа для 3D.

5. Результаты вычислительных экспериментов

Ниже приведены результаты статистического эксперимента по генерации графов первых ближайших соседей с помощью алгоритма, описанного выше в п.3. Статистика собиралась по миллиону генераций. Подчеркнем еще раз, что расчеты по сгенерированной, а не вычисленной матрице случайных расстояний проводились безотносительно размерности пространства вершин графа. Полученные результаты сравниваются затем с результатами генерации графов ближайших соседей в пространствах различной размерности, основанными на случайном задании координат точек и последующем вычислении расстояний между ними. На Рис. 2-5 показаны результаты расчета статистик графов для алгоритма генерации матрицы расстояний без учета размерности пространства (раздел 3).

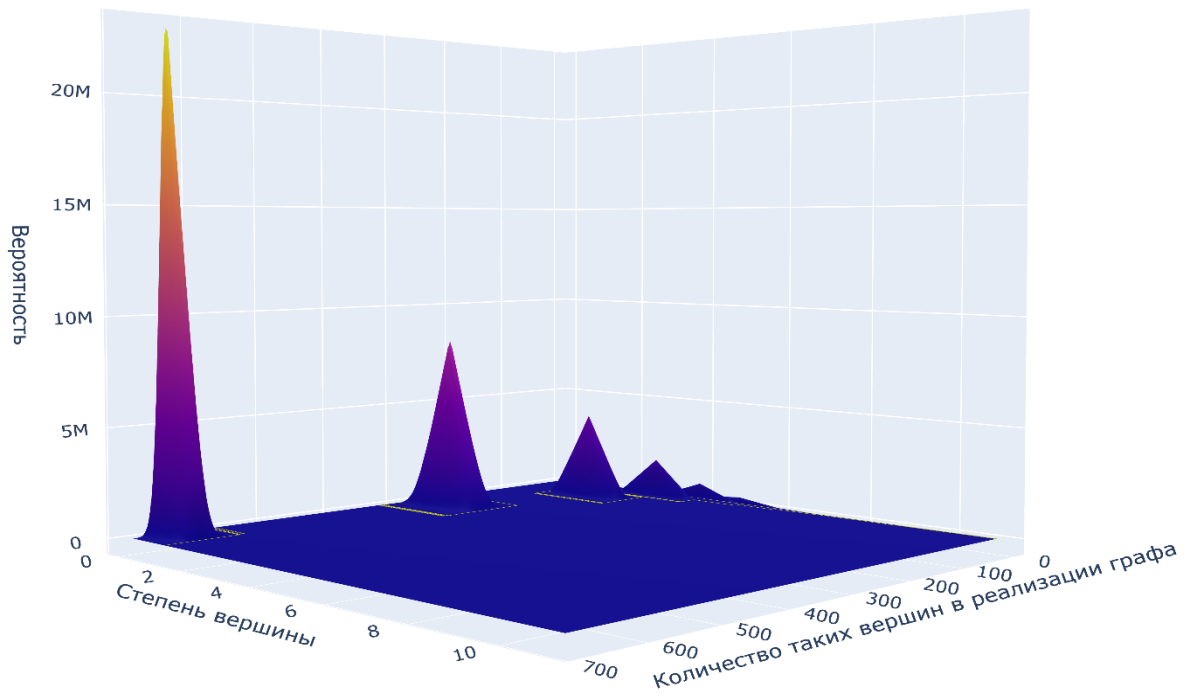


Рис. 2 – Распределение вершин первых соседей по степеням для $N = 1000$

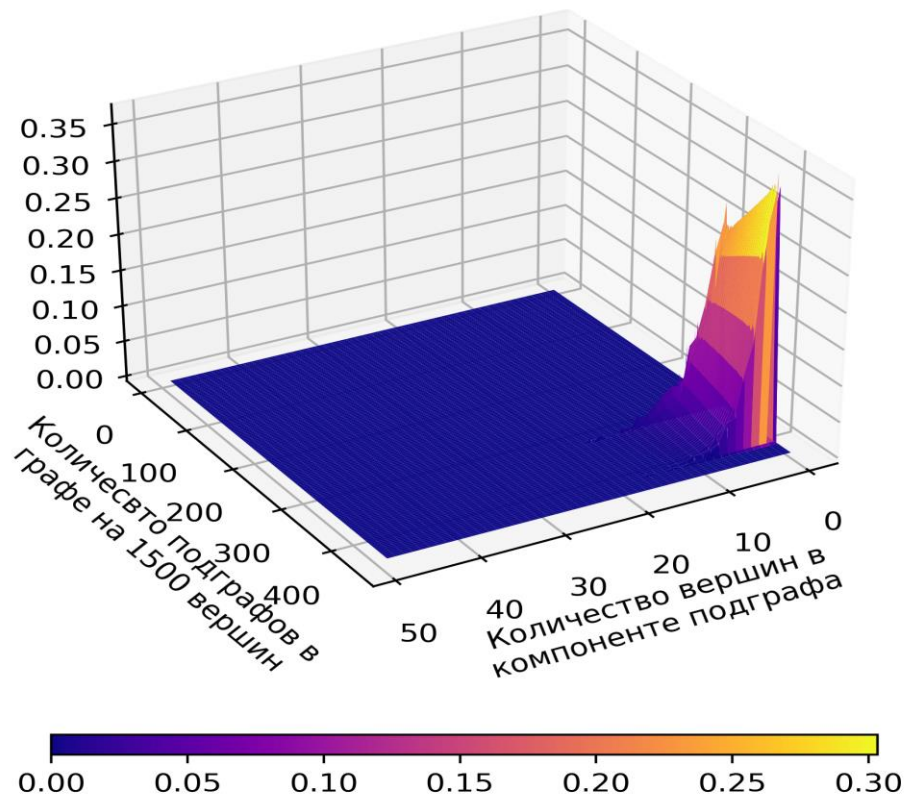


Рис. 3 – Совместное распределение графов 1-NN по числу фрагментов и по числу вершин для $N = 1500$

По итогам моделирования можно ответить на вопрос, какова вероятность $f_P(n)$ того, что случайный граф содержит данное число n вершин степени P по входящим ребрам. Например, распределение $f_0(n)$ для $P=0$ приведено на рис. 4 для $N=300$.

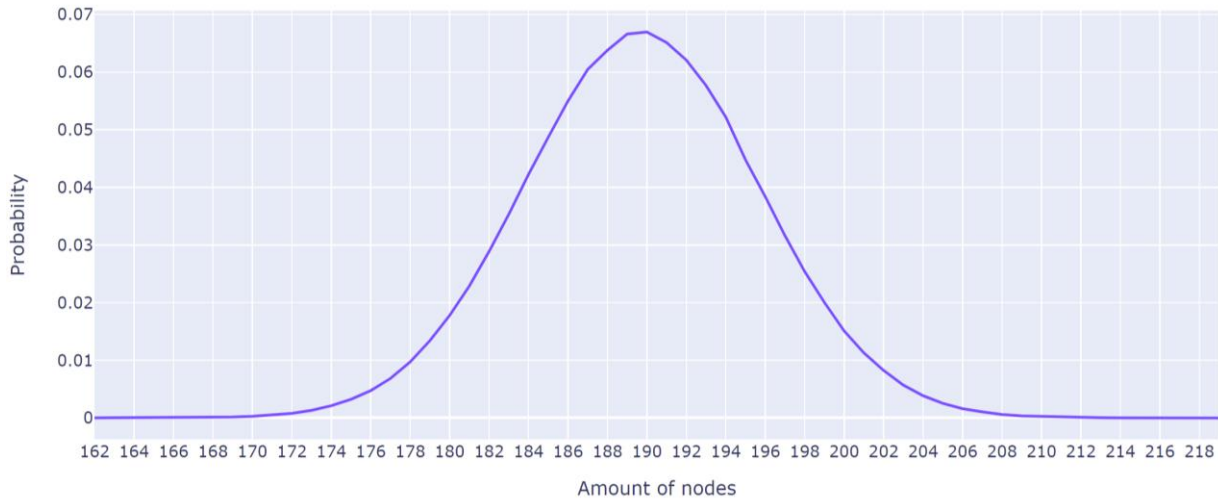


Рис. 4 – Распределение числа вершин со степенью $P=0$ для $N=300$

Аналогично можно построить распределения числа вершин с первой степенью, второй и т.д. Если расположить эти распределения на одной плоскости в виде «поверхности» $F(P,n)$, которая есть совокупность распределений $f_P(n)$, то получим фазовое пространство возможных реализаций графа. При этом возможна не любая траектория $(P, n(P))$, а лишь такая, для которой выполняются условия

$$\sum_P n(P) = N, \quad \sum_P Pn(P) = N. \quad (5.1)$$

Важным результатом численного анализа явилось то, что эмпирическое распределение $f_P(n)$ оказалось весьма близко к нормальному. Например, для плотностей распределения числа вершин нулевой и первой степеней в графе из N вершин имеем аппроксимацию с детерминацией выше 0,999:

$$f_0(n) \approx B_0 N \exp\left(-A_{00}(n - \mu_0 N)^2\right), \quad B_0 = 0,0667, \quad A_{00} = 0,014, \quad \mu_0 = 0,63. \quad (5.2)$$

$$f_1(n) \approx B_1 N \exp\left(-A_{11}(n - \mu_1 N)^2\right), \quad B_1 = 0,0587, \quad A_{11} = 0,010, \quad \mu_1 = 0,26. \quad (5.3)$$

Распределения остальных степеней имеют похожий вид. Однако в силу условия (5.1) эти распределения коррелированы. Например, совместное распределение числа вершин степеней 0 и 1 показано на рис. 5.

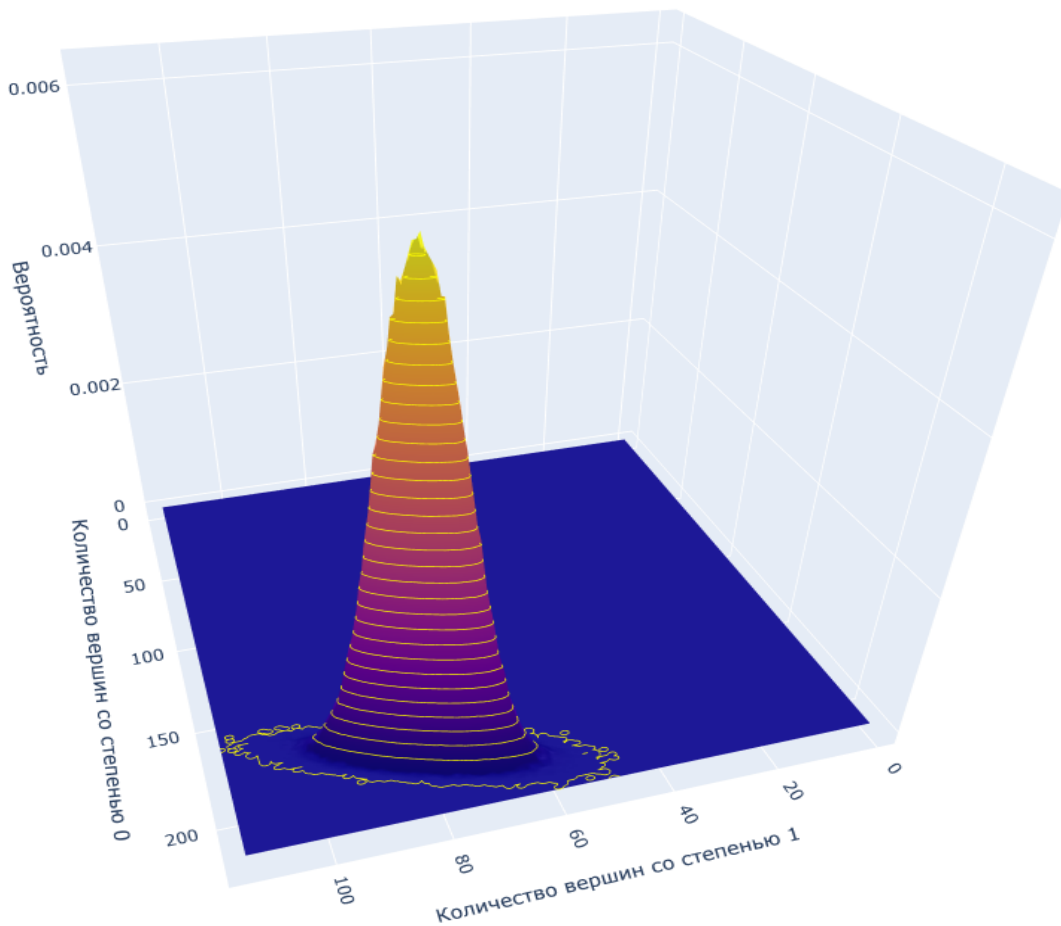


Рис. 5 – Совместное распределение числа вершин со степенями 0 и 1 для $N = 300$

Носитель этого распределения имеет вид вытянутого вдоль диагонали эллипса. Это означает, что эмпирическое совместное распределение числа вершин по степеням может быть приближенно представлено многомерным нормальным распределением:

$$f(n_0, n_1, \dots, n_p) = \frac{N}{(2\pi)^{p/2} \sqrt{\det C}} \exp \left(-\frac{1}{2} \sum_{i,j=0}^p C_{ij}^{-1} (n_i - N\mu_i)(n_j - N\mu_j) \right), \quad (5.4)$$

где $\det C$ есть определитель ковариационной матрицы, а C_{ij}^{-1} – элементы обратной ковариационной матрицы. Приближенность формулы (5.4) состоит в том, что аргументы функции f в (5.4) неотрицательны, а нормировка отвечает интегралам от $-\infty$ до $+\infty$. Такое приближение может быть использовано по причине быстрого стремления плотности распределения к нулю при уменьшении числа вершин (см. Рис. 4).

То, что распределение степеней вершин действительно является гауссовым, следует из локальной предельной теоремы Муавра-Лапласа, примененной к полиномиальному распределению вершин по степеням. Действительно, если μ_j есть вероятность того, что вершина графа имеет степень j , то вероятность того, что n_j вершин из N имеют степень j , равна

$C_N^{n_j} (\mu_j)^{n_j} (1 - \mu_j)^{N - n_j}$, что в пределе больших N имеет гауссову асимптотику. Так что распределение вида (5.4) вполне ожидаемо для коррелированных нормально распределенных случайных величин и подтверждает корректность проведенного вычислительного эксперимента. Нетривиальным же вычислительным фактом оказалось то, что вероятности μ_j , фигурирующие в (5.4), сами имеют нормальное распределение для достаточно больших значений N . По результатам анализа графов с числом вершин N от 300 до 1500 на основе миллиона экспериментов для каждого варианта числа вершин были получены следующие значения μ_j , не зависящие от N :

$$\mu_0 = 0,632; \mu_1 = 0,264; \mu_2 = 0,080; \mu_3 = 0,019;$$

$$\mu_4 = 0,0036; \mu_5 = 0,00059; \mu_6 = 0,00008.$$

Эти величины имеют гауссову аппроксимацию с детерминацией 0,999:

$$\mu_j \approx 3,7268 \exp(-0,1074(j + 4,01)^2), \quad j = 0, 1, \dots \quad (5.5)$$

В результате получается, что согласована как нормировка распределения (5.4), так и сумма средних значений степеней, равная N .

В качестве примера приведем конкретные значения прямой и обратной ковариационных матриц для совместного распределения степеней 0, 1, 2 и 3 в (5.4):

$$C_{ij} = \begin{bmatrix} 87,14 & -80,72 & -8,28 & 1,03 \\ -80,72 & 119,10 & -29,15 & -7,66 \\ -8,28 & -29,15 & 43,03 & -4,45 \\ 1,03 & -7,66 & -4,45 & 11,75 \end{bmatrix}, \quad C_{ij}^{-1} = \begin{bmatrix} 0,425 & 0,410 & 0,399 & 0,381 \\ 0,410 & 0,407 & 0,394 & 0,379 \\ 0,399 & 0,394 & 0,406 & 0,376 \\ 0,381 & 0,379 & 0,376 & 0,441 \end{bmatrix}$$

Определитель самой ковариационной матрицы в данном примере приблизительно равен $\det C = 10^5$ (точнее, 102019,23).

Рассмотрим теперь статистику структур графов первых соседей для случая, когда координаты точек имеют равномерное распределение в единичном n -мерном кубе (метрика евклидова).

В одномерном случае картина в определенном смысле вырожденная, поскольку ближайших соседей у каждой точки либо один, либо два. Возможные степени вершин по входящим ребрам – ноль, один и два. При этом доля вершин со степенями 0 и 2 равна отношению числа фрагментов, из которых состоит граф из N вершин, к числу этих вершин, поскольку каждый фрагмент начинается вершиной со степенью 0 и заканчивается циклом длины 2, и только одна вершина, входящая в этот цикл, имеет степень 2. То есть число вершин со степенью 0 и 2 в каждой реализации графа из N вершин равно числу возникающих при этом несвязных фрагментов. В итоге получаем распределения, представленные на Рис. 6-8. Статистика получена генерацией графа из $N=1000$ вершин, число генераций – 1 млн. Такой же численный эксперимент был затем поставлен и для размерностей 2, 3, 4 и 5. Нижеследующие графики представляют бенчмарк статистик для графов первых ближайших соседей в пространствах указанных размерностей.

В размерности 1D более 7 вершин в несвязном фрагменте графа на 1000 вершин практически не встречается, наиболее вероятное число фрагментов 335.

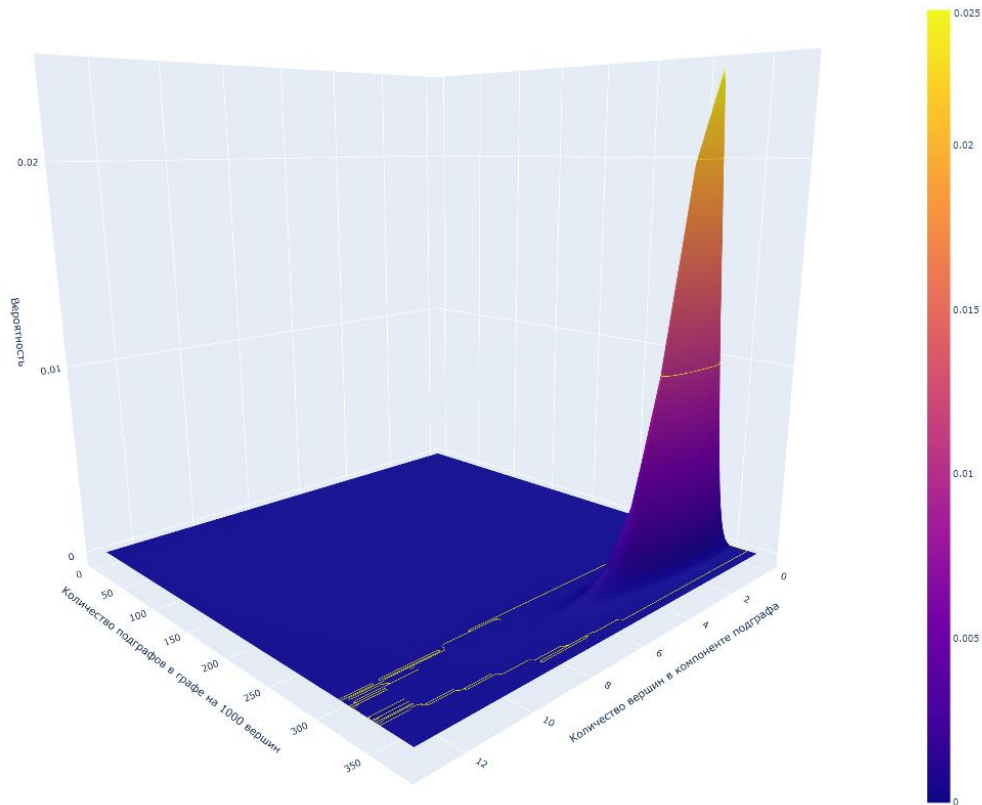


Рис. 6 – Распределение графов 1-NN для 1D-соседей по числу фрагментов и числу вершин во фрагменте

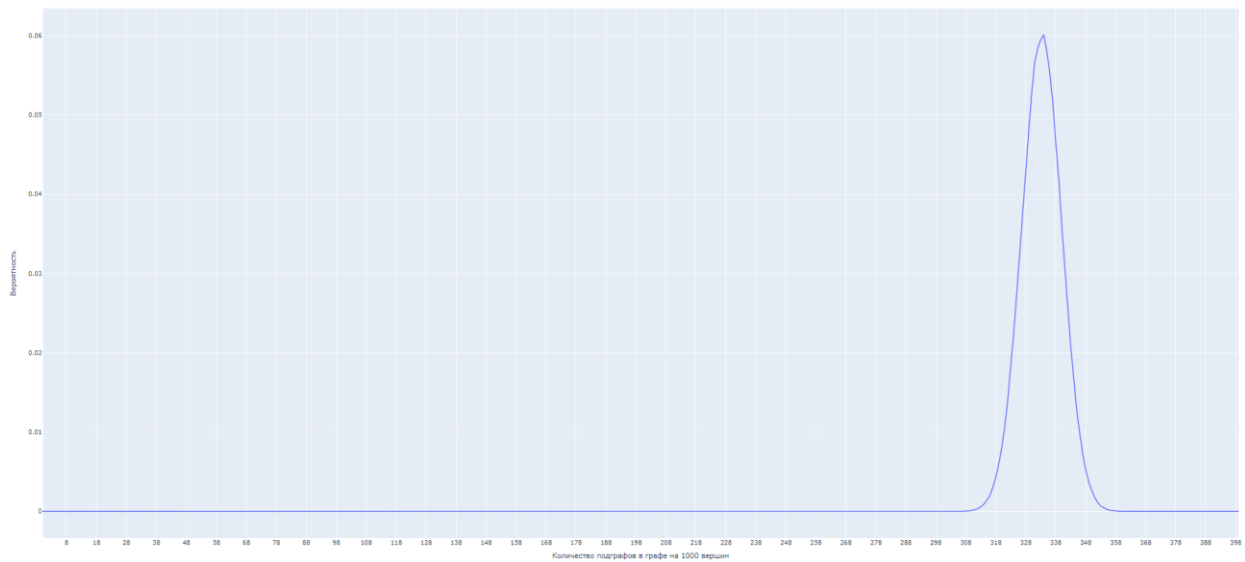


Рис. 7 – Распределение графов 1-NN для 1D-соседей по числу фрагментов

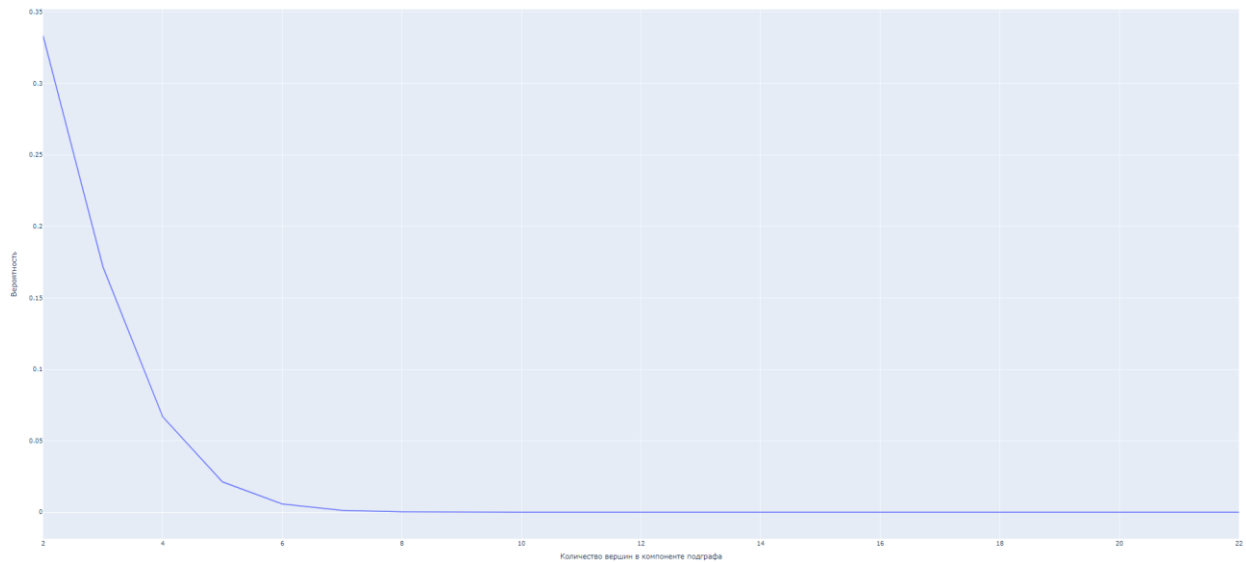


Рис. 8 – Распределение графов 1-NN для 1D-соседей по числу вершин во фрагменте

Рассмотрим теперь двумерный случай. Совместное распределение структур графов 1-NN по числу фрагментов и числу вершин во фрагментах представлено на рис. 9.

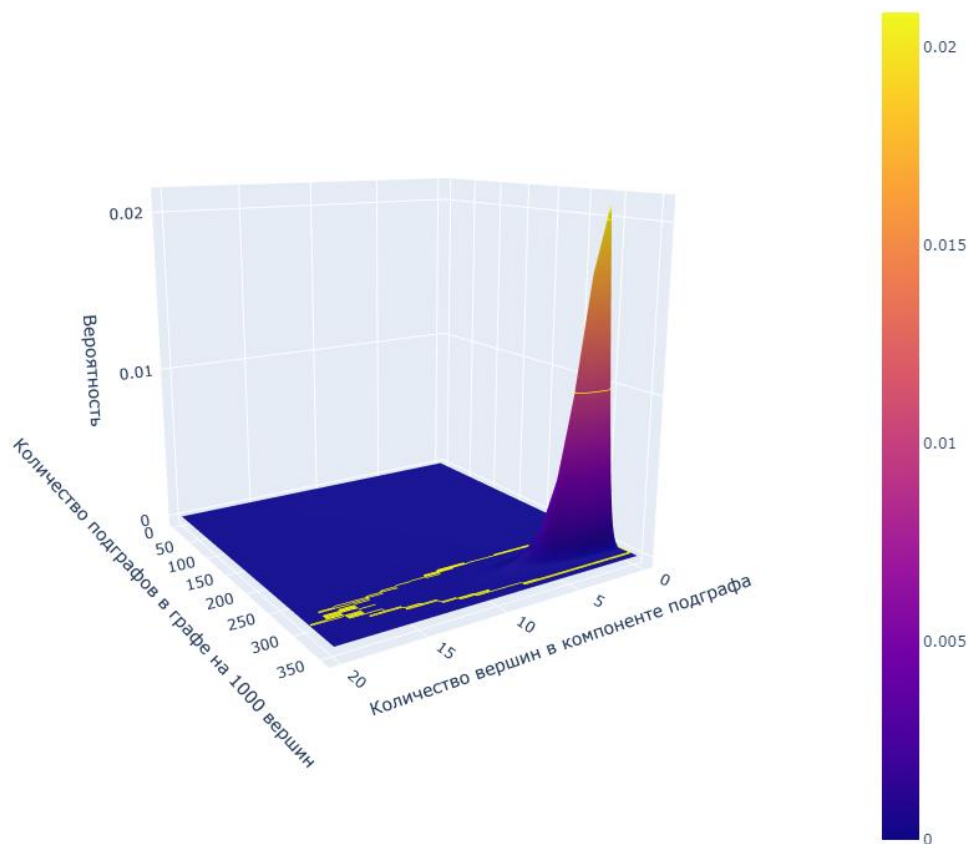


Рис. 9 – Распределение графов 1-NN для 2D-соседей по числу фрагментов и числу вершин во фрагменте

Как на Рис. 6, так и на Рис. 9 наблюдается некоторое отличие от аналогичного «безразмерного» распределения на Рис. 3, в котором вершина (максимум вероятности) отвечает количеству подграфов, равному $N/4$. Здесь же максимум смещен вправо и отвечает примерно значению 310 вместо ожидаемых 250 для распределения Вигнера. Однако по сравнению с одномерным случаем вершина распределения в размерности 2D смещена влево, то есть находится несколько ближе к значению $N/4$.

Эмпирическое распределение только количества подграфов в графе на 1000 вершин показано на рис. 10.

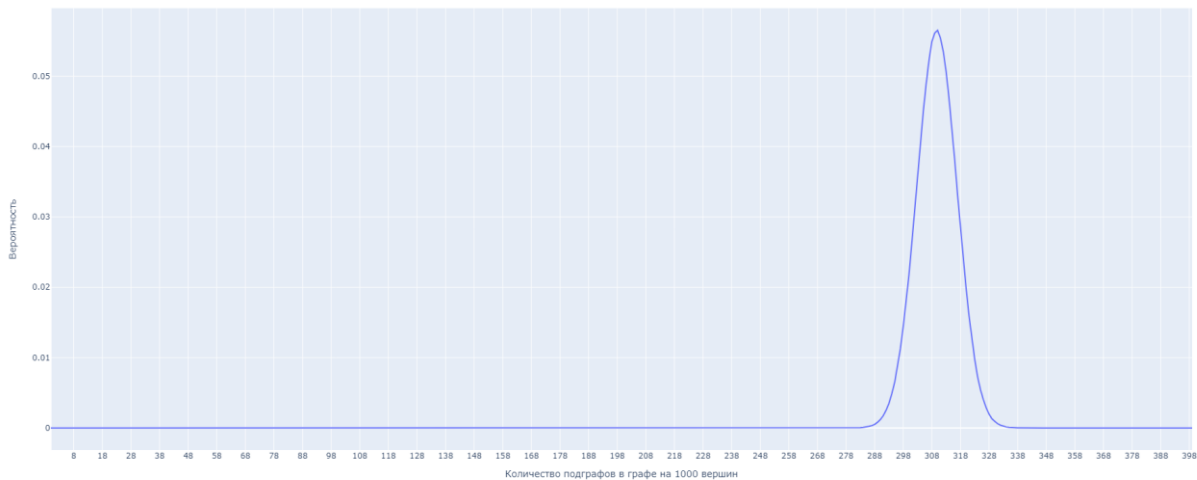


Рис. 10 – Распределение графов 1-NN для 2D-соседей по числу фрагментов

На рис. 11 показан другой аспект совместного распределения – по числу вершин во фрагменте.

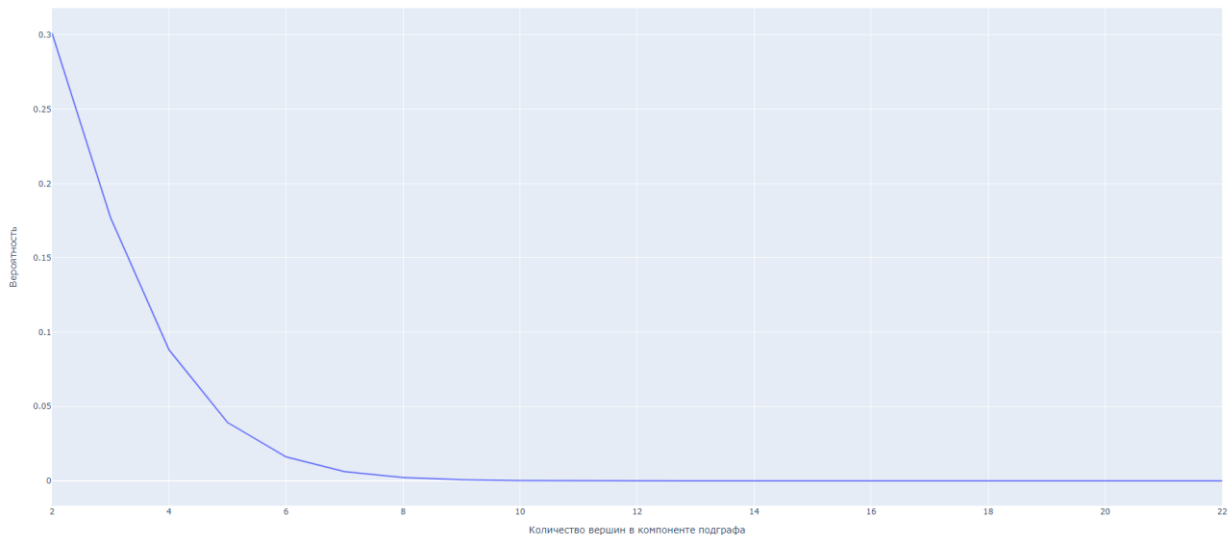


Рис. 11 – Распределение графов 1-NN для 2D-соседей по числу вершин во фрагменте

Наиболее вероятным числом вершин является 2, то есть большая часть фрагментов представляет собой изолированные циклы длины 2. Число вершин во фрагменте, большее 8, наблюдается уже крайне редко.

На рис. 12 представлено совместное распределение графа по числу вершин и вершин по степеням.

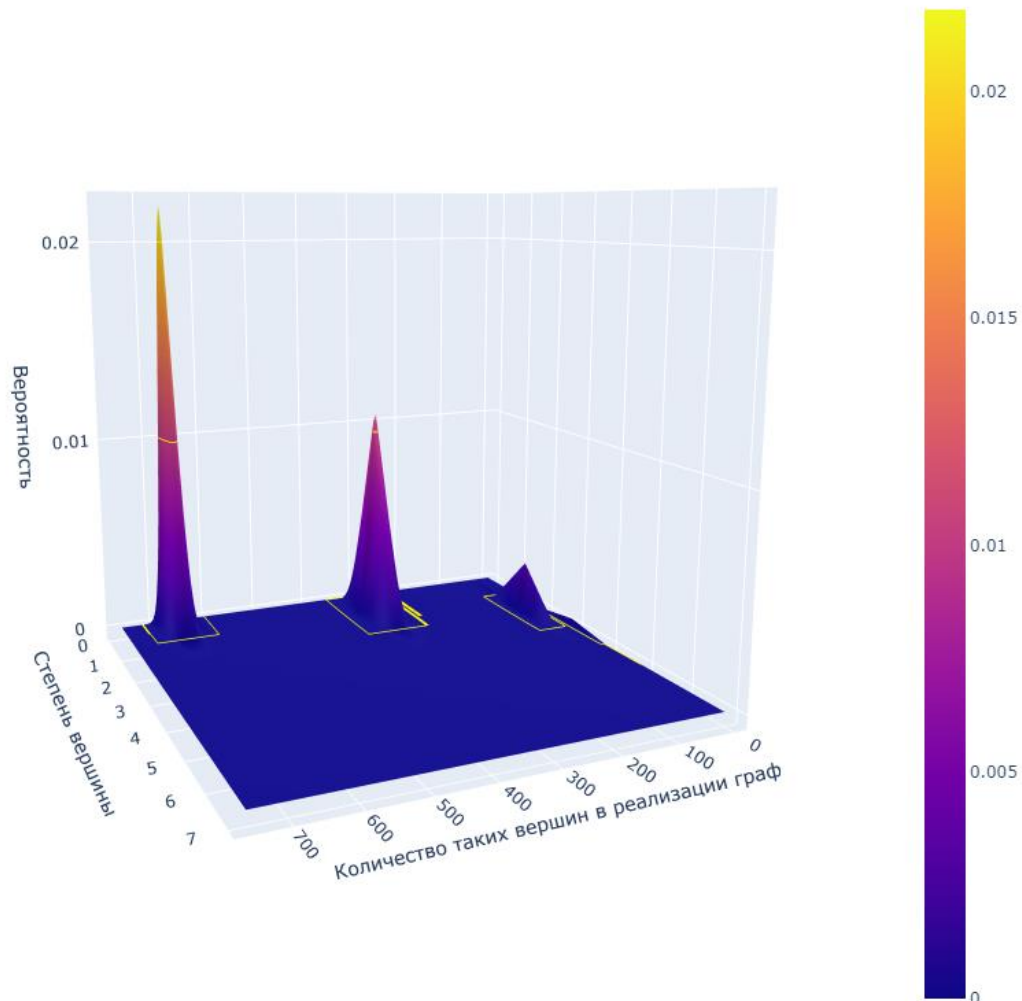


Рис. 12 – Распределение вершин графов 1-NN для 2D-соседей по степеням

Выяснилось, что доля вершин со степенью 0 составила приблизительно величину 0,662, со степенью 1 – величину 0,299, со степенью 2 – величину 0,039, а со степенью 3 – величину 0,0008. Также встретились вершины со степенью 4, их количество составило 300 на миллион генераций графов на 1000 вершин (т.е. полное число сгенерированных вершин в численном эксперименте составило 10^9). Отметим, что максимально возможное число первых ближайших соседей, равное в данном случае пяти, не встретилось ни разу. Это связано с тем, что такая ситуация должна быть специально устроена: все пять расстояний от некоторых случайных точек до некоторой шестой точки должны быть равны между собой, что при генерации случайных координат технически не реализуется – там все случайные расстояния различны.

В трехмерном пространстве соответствующие распределения структур графов первых ближайших соседей представлены на рис. 13-16. Как и ранее, генерировались структуры на 1000 вершин, число генераций – 1 млн.

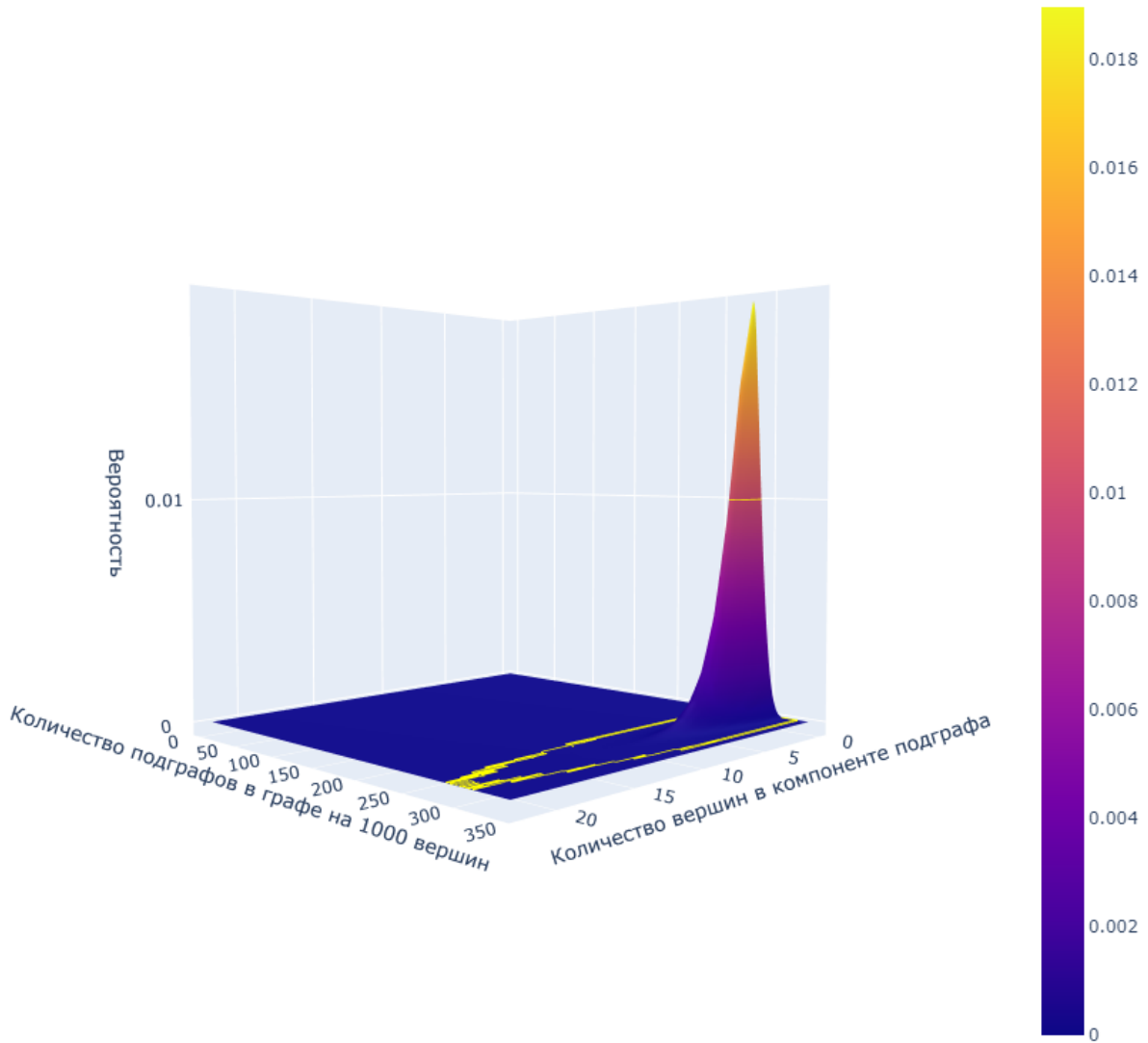


Рис. 13 – Распределение графов 1-NN для 3D-соседей по числу фрагментов и числу вершин во фрагменте

Сравнивая распределение на рис. 13 с аналогичным распределением на рис. 3, видим, что максимум вероятности заметно сместился влево: наиболее вероятное число фрагментов равно примерно 290. Возникает предположение, которое подтверждается дальнейшими вычислениями, что для случая 5D точка максимума будет близка к 250, т.е. к значению для распределения, не зависящего от размерности.

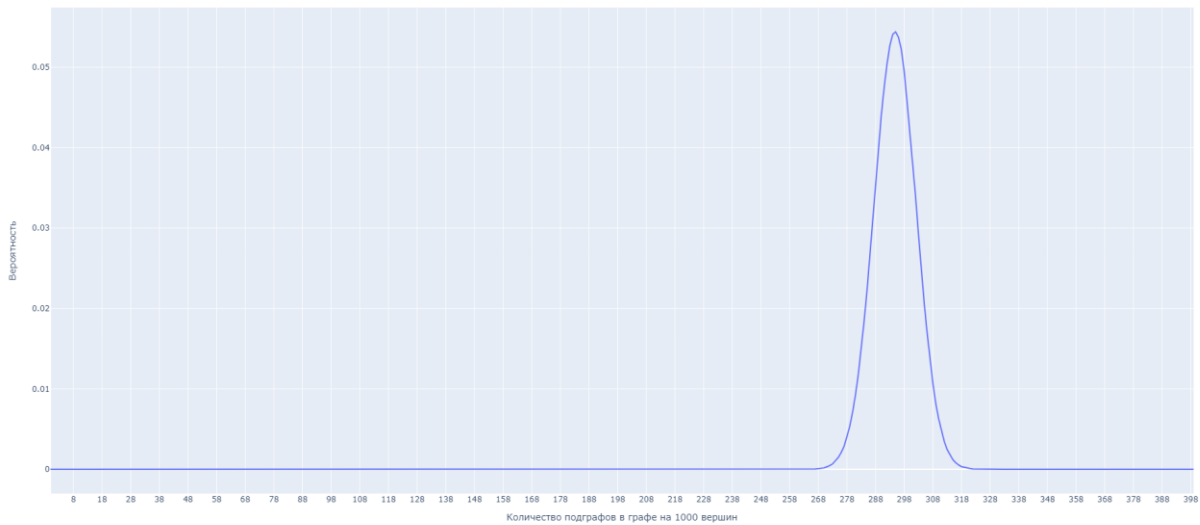


Рис. 14 – Распределение графов 1-NN для 3D-соседей по числу фрагментов

Число возможных вершин во фрагменте для случая 3D увеличилось по сравнению с 2D и стало равным примерно 10, более крупные фрагменты очень маловероятны.

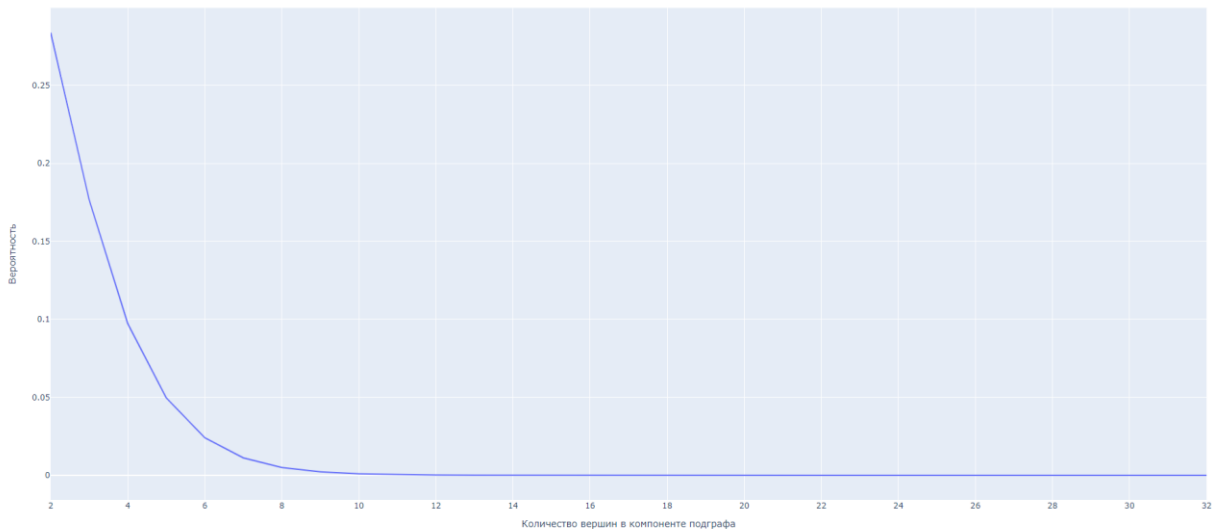


Рис. 15 – Распределение графов 1-NN для 3D-соседей по числу вершин во фрагменте

Распределение по степеням и числу вершин представлено на рис. 16. Интересно проследить за изменением долей вершин с заданными степенями, то есть за величинами μ_j в терминах формулы (5.5). Выяснилось, что теперь доля вершин со степенью 0 составила приблизительно величину 0,652, со степенью 1 – величину 0,291, со степенью 2 – величину 0,053, а со степенью 3 – величину 0,004. Эти значения приблизились к величинам (5.5). Также встретились вершины со степенями 4, 5 и 6. Их доли равны соответственно 10^{-4} , 10^{-6} и 10^{-9} .

Как и в примерах ранее, полное число сгенерированных вершин в этом эксперименте составило 10^9 .

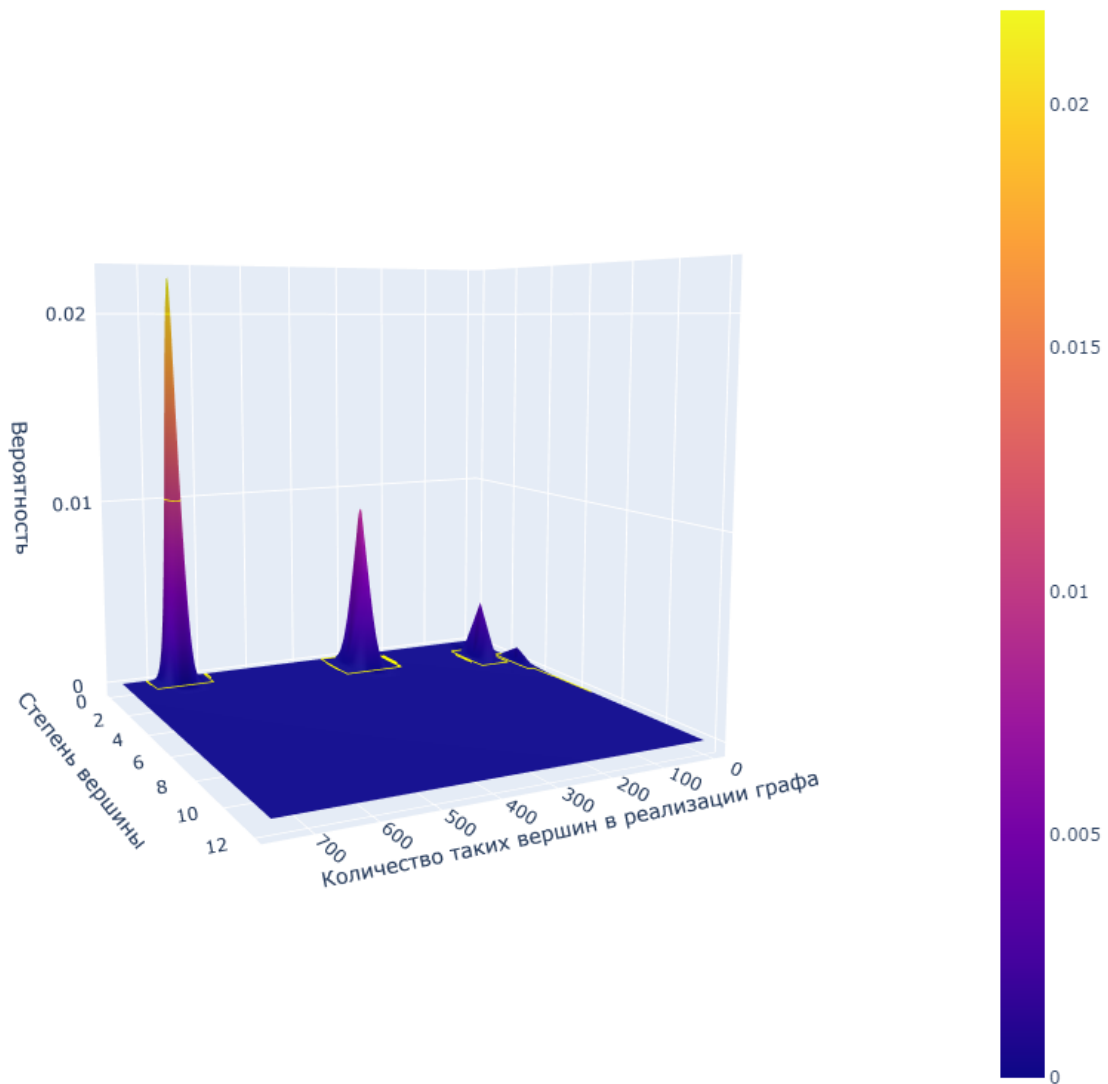


Рис. 16 – Распределение вершин графов 1-NN для 3D-соседей по степеням

Далее приведем результаты расчетов для размерности 4D. Отмеченные выше тенденции сохраняются: увеличивается число вершин во фрагменте, максимум распределения сдвигается влево, уменьшается доля коротких фрагментов из двух и трех вершин, увеличивается наблюдаемая максимальная степень ближайшего соседства – теперь она равна 8. Соответствующие распределения приведены на Рис. 17-20. При этом доля вершин со степенью 0 составила приблизительно величину 0,645, со степенью 1 – величину 0,286, со степенью 2 – величину 0,061, а со степенью 3 – величину 0,007.

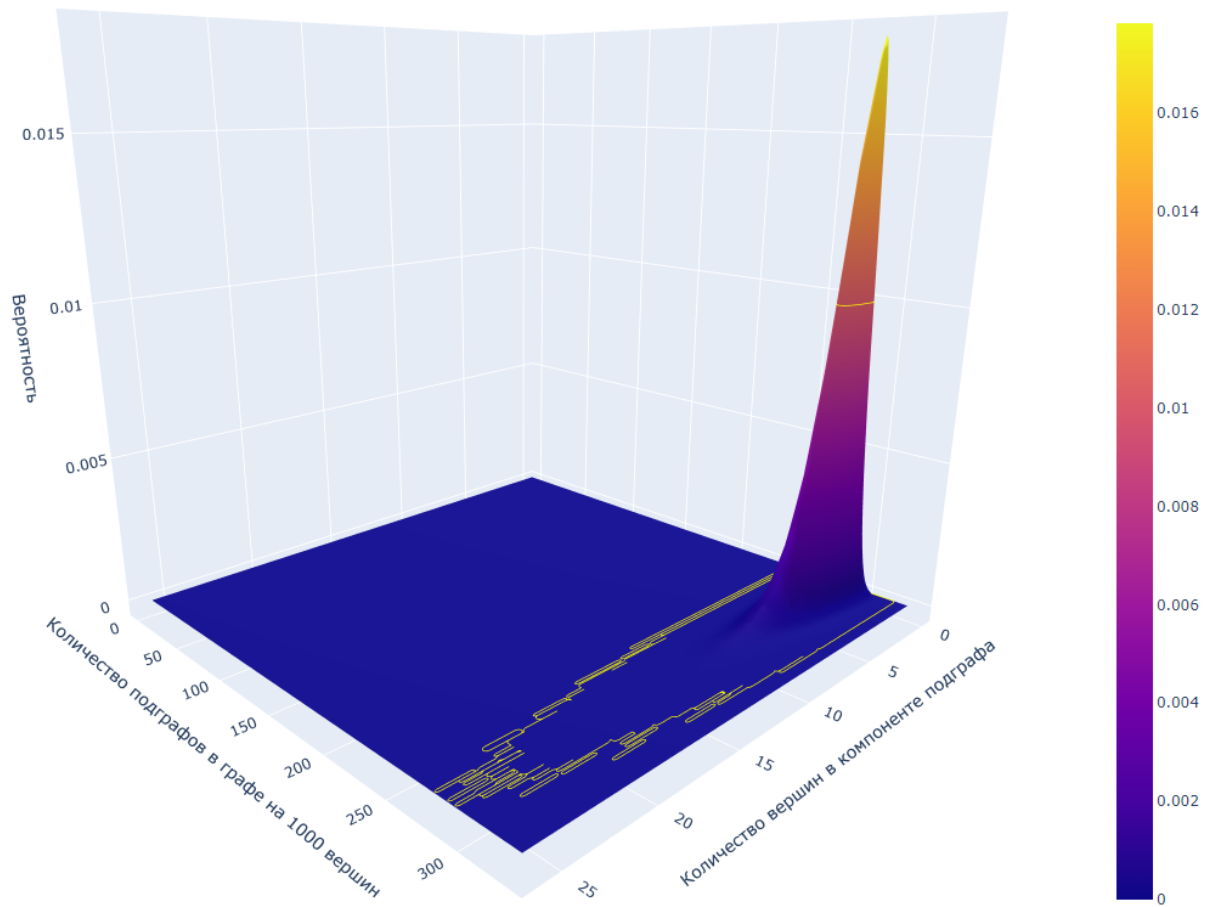


Рис. 17 – Распределение графов 1-NN для 4D-соседей по числу фрагментов и числу вершин во фрагменте

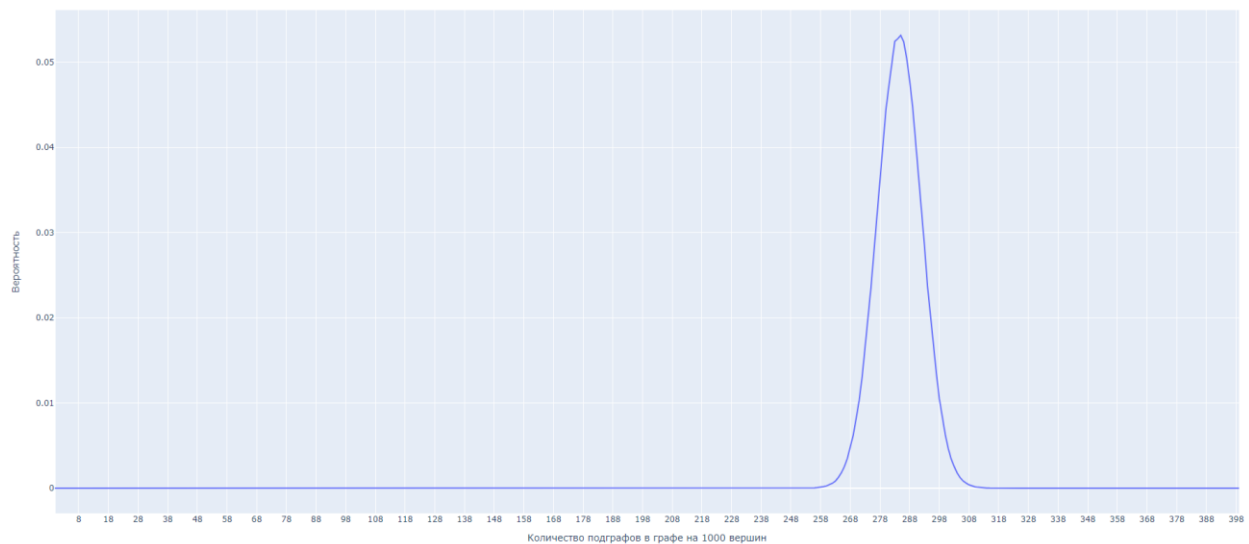


Рис. 18 – Распределение графов 1-NN для 4D-соседей по числу фрагментов

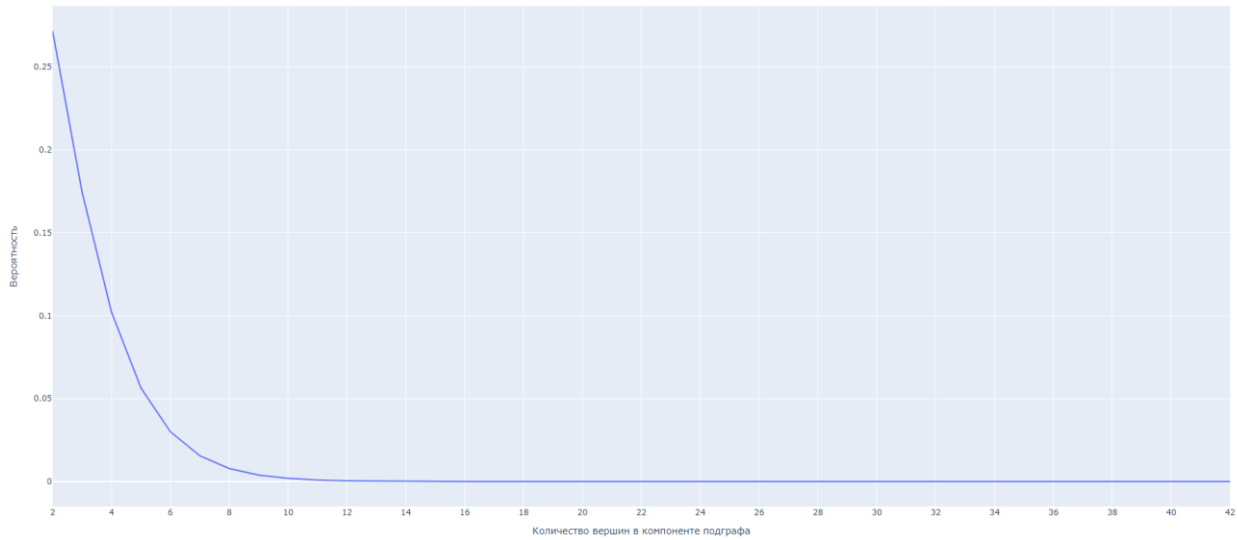


Рис. 19 – Распределение графов 1-NN для 4D-соседей по числу вершин во фрагменте

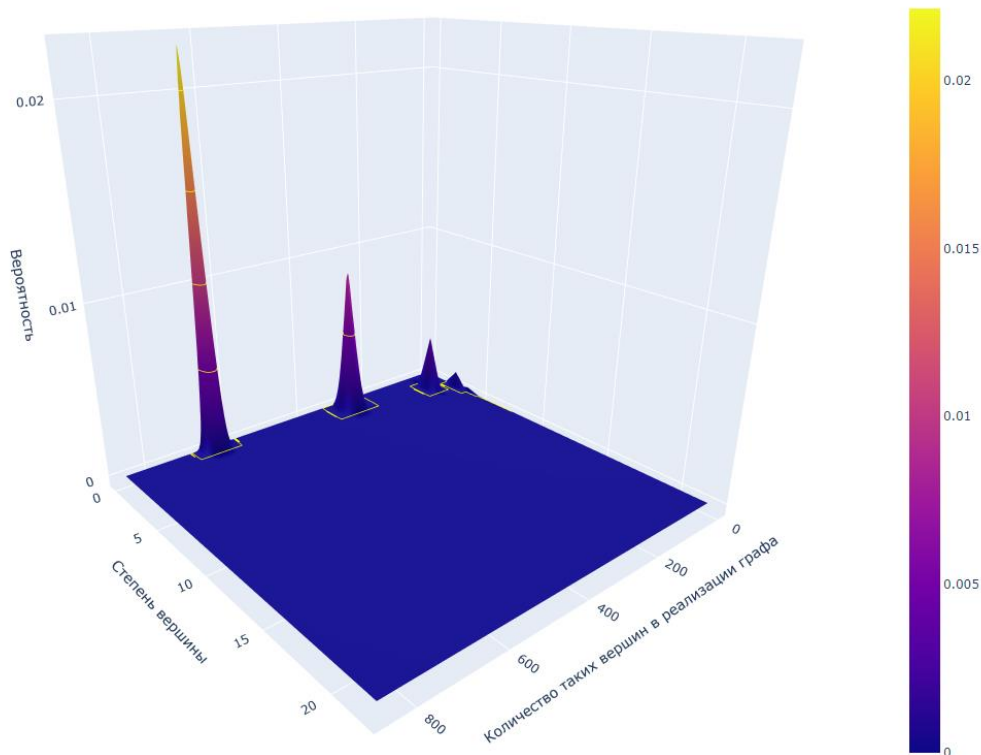


Рис. 20 – Распределение вершин графов 1-NN для 4D-соседей по степеням

Аналогичные распределения получаются и для размерности 5D (рис. 21-24). При этом максимум распределения графов по числу фрагментов для 1000 вершин достигается в точке 265, что довольно близко к «безразмерному» распределению. Для распределения степеней вершин первых соседей в 5D получаем следующие эмпирические оценки:

$$\mu_0^{(5D)} = 0,639; \mu_1^{(5D)} = 0,283; \mu_2^{(5D)} = 0,067; \mu_3^{(5D)} = 0,009; \mu_4^{(5D)} = 0,0008. \quad (5.6)$$

Центр эмпирического распределения на Рис. 17 находится в точке 283, на Рис. 21 – в точке 275, т.е. он действительно смещается в сторону центра распределения Вигнера. В размерности 5D появились также две вершины со степенью 9.

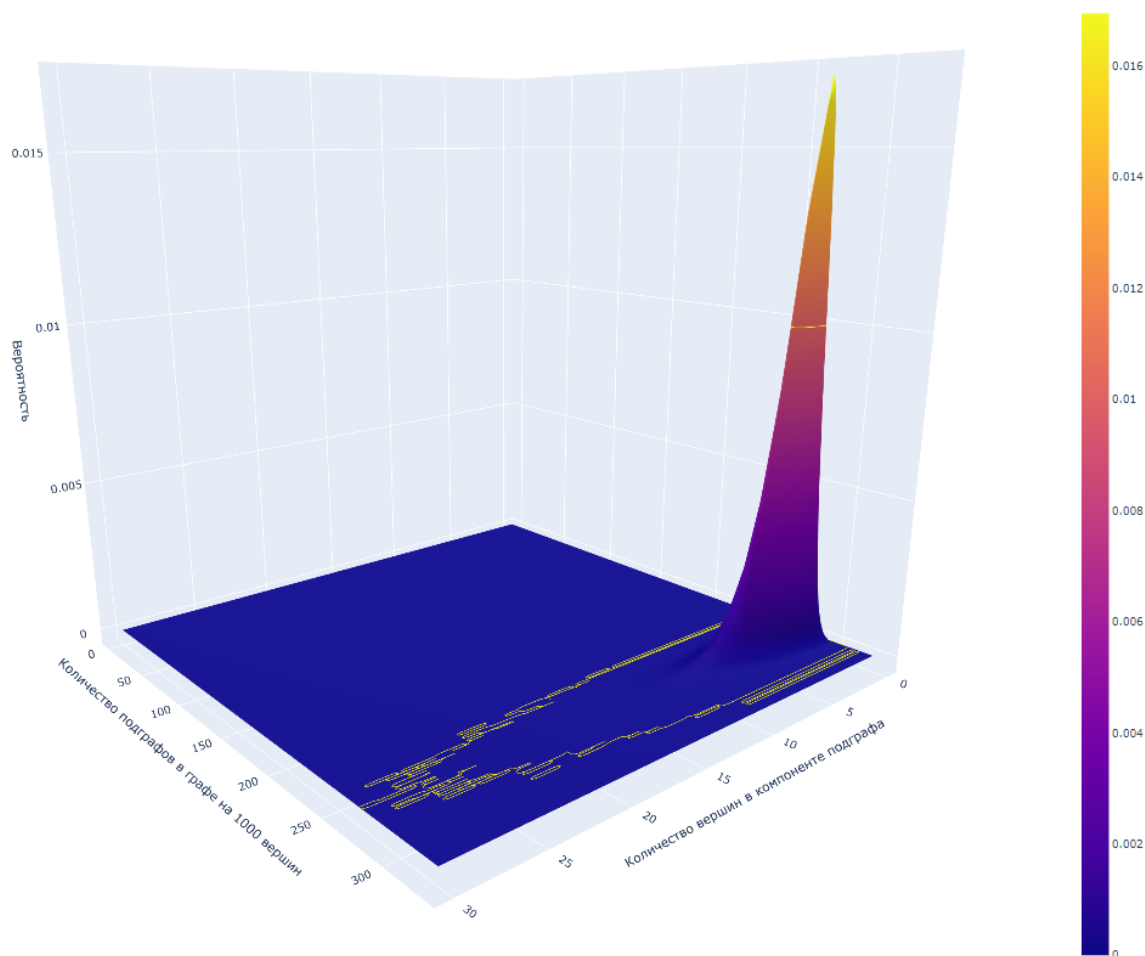


Рис. 21 – Распределение графов 1-NN для 5D-соседей по числу фрагментов и числу вершин во фрагменте

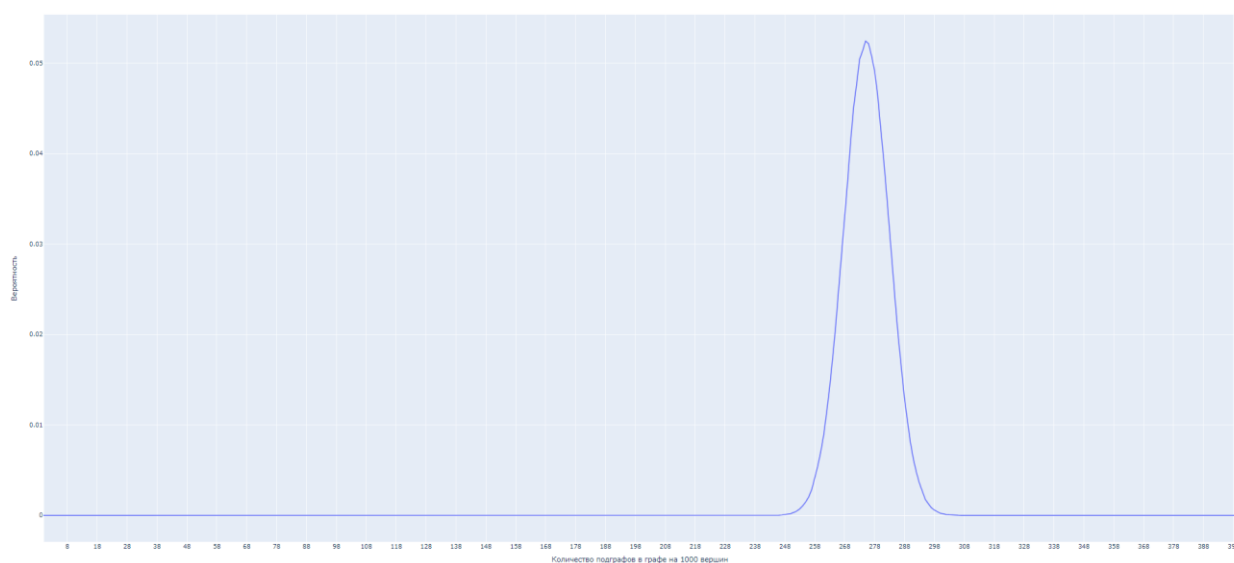


Рис. 22 – Распределение графов 1-NN для 5D-соседей по числу фрагментов

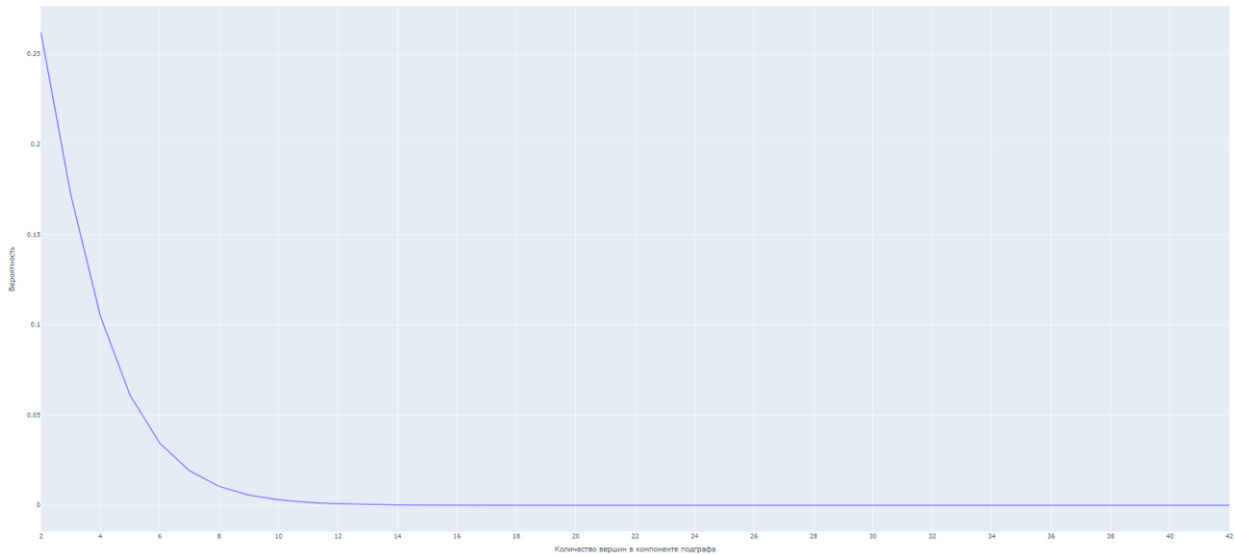


Рис. 23 – Распределение графов 1-NN для 5D-соседей по числу вершин во фрагменте

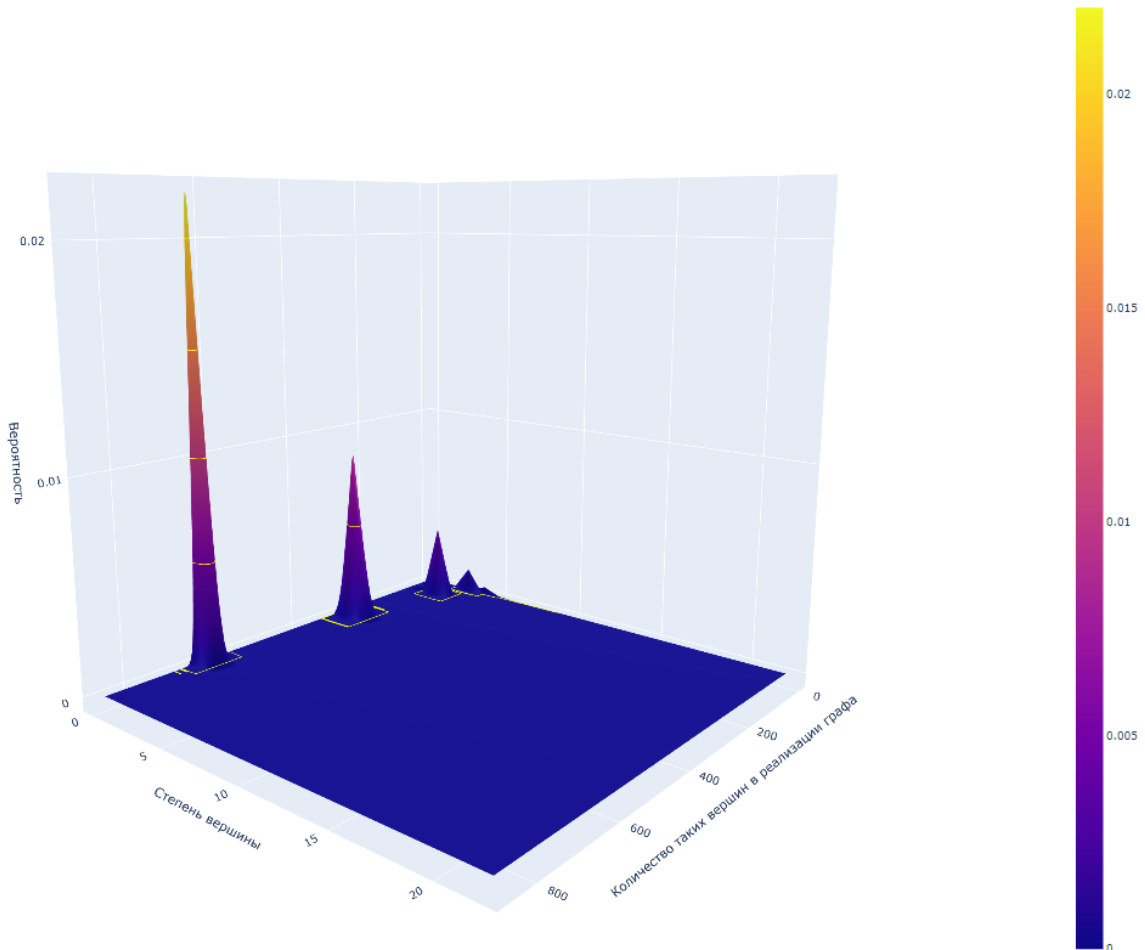


Рис. 24 – Распределение вершин графов 1-NN для 5D-соседей по степеням

Возникает вопрос: насколько близки между собой распределения вероятностей μ_j из (5.5) и аналогичных величин в примере с размерностью 5D. Соответствующие распределения представлены на рис. 25. Если близость достаточна с практической точки, то вместо трудоемкой генерации структур

графов в некоторой произвольной размерности можно было бы использовать универсальный безразмерный бенчмарк, построенный выше.

Большое число сгенерированных выборок приводит к тому, что распределения, представленные на рис. 25, четко опознаются как различные. Это естественно, поскольку распределения сближаются не при увеличении числа генераций или числа вершин графа, а при увеличении размерности пространства, в котором расположены точки. Однако вопрос в данном случае не в том, чтобы проверить, на каком уровне значимости совпадают кривые на рис. 25, а в том, насколько изменится уровень значимости при проверке статистической гипотезы для некоторого конкретного графа, если в критерии сравнения заменить одно распределение на другое.

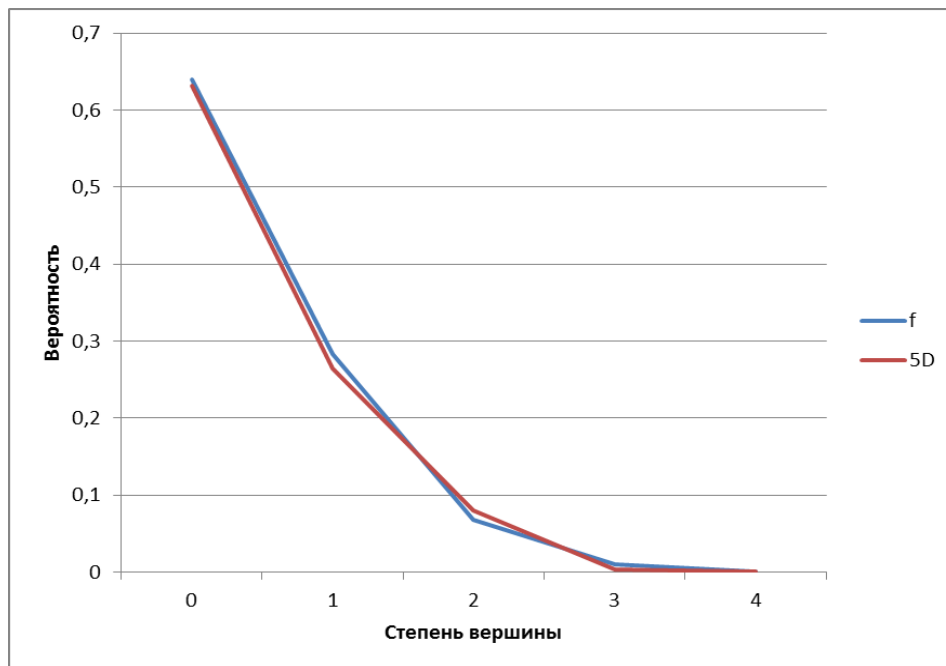


Рис. 25 – Сравнение «безразмерного» распределения вероятностей вершин по степеням и распределения, отвечающего размерности 5D

Рассмотрим, например, ситуацию, когда для некоторого графа с достаточно большим числом N вершин, координаты которых расположены в единичном кубе размерности $n=5$, получено некоторое распределение вероятностей $\tilde{f}(j)$ вершин по степеням. Мы хотим оценить вероятность реализации этой структуры или, как вариант, оценить уровень значимости, на котором это распределение $\tilde{f}(j)$ можно считать выборочным из распределения $f^{(5)}(j)$, используя для этой цели «безразмерное» распределение $f(j)$, $j=0,1,\dots,p$. Поскольку реально степеней вершин у нас не много, можно применить критерий согласия хи-квадрат (см. [92], с. 204). Для этого надо вычислить статистику

$$\chi^2 = N \sum_{j=0}^p \frac{(\tilde{f}(j) - f^{(5)}(j))^2}{f^{(5)}(j)}. \quad (5.7)$$

Если $\chi^2 < \chi_\alpha^2(p)$, где p есть число степеней свободы распределения хи-квадрат, равное в данном случае максимальной степени вершины в эмпирическом распределении, то на уровне значимости α гипотеза о равенстве распределений принимается.

Рассмотрим величину

$$u = \frac{\chi^2}{N} + 1 = \sum_{j=0}^p \tilde{f}^2(j) / f^{(5)}(j). \quad (5.8)$$

При замене $f^{(5)}(j)$ на $f(j)$ в критерии (5.7) получим

$$\chi'^2 = N \sum_{j=0}^{\infty} \frac{(\tilde{f}(j) - f(j))^2}{f(j)}. \quad (5.9)$$

Поскольку же $\tilde{f}(j) = 0$ при $j > p$, то

$$\frac{|\chi'^2 - \chi^2|}{N} = |u' - u| \leq \sum_{j=0}^p \tilde{f}^2(j) \frac{|f^{(5)}(j) - f(j)|}{f^{(5)}(j)f(j)}. \quad (5.10)$$

Сделаем оценку для равномерного эмпирического распределения $\tilde{f}(j) = 1/(p+1)$. По имеющемуся бенчмарку значений для $f^{(5)}(j)$ и $f(j)$ находим, что

$$\sum_{j=0}^4 \frac{|f^{(5)}(j) - f(j)|}{f^{(5)}(j)f(j)} \approx 591, \quad \sum_{j=0}^4 \frac{1}{f^{(5)}(j)} \approx 1962,$$

так что средневзвешенное по обратным частотам относительное изменение частоты составляет величину примерно 0,3. Это есть характерная величина относительного изменения статистики u из (5.8). Теперь надо определить, как меняется уровень значимости критерия $\chi_\alpha^2(4)$ при относительном изменении самой статистики на 0,3. Чувствительность уровня значимости зависит от самого уровня значимости, и в данном случае имеет следующие значения (таблица 2). Положим $\delta(\alpha) = \frac{\alpha(\chi) - \alpha(\chi')}{1 - (\chi/\chi')^2}$. Тогда по таблице процентных значений критерия хи-квадрат вычисляем:

Таблица 2. К оценке чувствительности уровня значимости

α	0,001	0,005	0,01	0,025	0,05	0,1	0,2	0,3	0,4	0,5
$\delta(\alpha)$	0,007	0,02	0,05	0,09	0,17	0,28	0,43	0,54	0,58	0,59

Чтобы вычислить, насколько изменится уровень значимости при вариации значения χ^2 , следует последнюю строку таблицы 2 умножить на относительное изменение χ^2 , в нашем случае это число равно 0,3. В результате получаем, что изменение уровня значимости с практической точки зрения

будет не очень значительным: вместо 0,001 мы получим 0,003, вместо 0,005 получим 0,01, вместо 0,01 получим 0,025, а вместо 0,25 получим 0,05.

Следовательно, распределения структур графов 1-NN в пространстве размерности пять и выше с достаточной точностью могут быть заменены распределениями, которые построены непосредственно генерацией расстояний между точками безотносительно размерности пространства. Это наблюдение обосновывает возможность использования эмпирически построенного бенчмарка для анализа графов в пространствах любых размерностей, больших четырех. Для первых же четырех размерностей используются данные, полученные в настоящем параграфе.

6. Заключение

В работе описаны вычислительные методы для генерации данных, необходимых для анализа и визуализации статистик структур графов ближайших соседей. Центральный результат состоит в том, что распределения графов по числу несвязных фрагментов, числу вершин во фрагменте и вершин по степеням, получаемые для различных размерностей, практически не зависят от предельных значений возможных ближайших соседей в зависимости от размерности пространства точек. В частности, статистики, получаемые для ближайших соседей в 5-мерном пространстве, оказываются уже очень близки к аналогичным результатам для способа генерации, не зависящим от размерности. Важность этого результата состоит в том, что это позволяет использовать для анализа графов в многомерном пространстве гораздо более экономный вычислительный алгоритм, получая результаты приблизительно на том же уровне значимости.

Разделы 2, 3, 5 написаны А.А. Кислицыным, раздел 4 написан М.В. Гогуевым, введение и заключение написаны совместно.

Литература

1. Кислицын А.А. Орлов Ю.Н. Исследование статистик графов ближайших соседей // Препринты ИПМ им. М.В. Келдыша. 2021. № 85. 23 с.
2. Кислицын А.А. Исследование статистик графов ближайших соседей // Математическое моделирование, 2022. Т. 34. № 8. С. 110-126.
3. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. – Springer, 2001. – 764 p.
4. Erdos P., Renyi A. On random graphs I // Publ. Math. Debrecen, 1959. V. 6. P. 290-297.
5. Колчин В.Ф. Случайные графы. – М.: Физматлит, 2004. – 256 с.
6. Bollobas B. Random Graphs. – Cambridge University Press, 2001. – 520 p.
7. Janson S., Luczak T., Rucinski A. Random graphs. – New York: Wiley, 2000. – 333 p.

8. Полянин А.Д., Манжиров А.В. Справочник по интегральным уравнениям. – М.: Физматлит, 2003. – 608 с.
9. Кендалл М., Моран П. Геометрические вероятности. – М.: Наука, 1972. – 192 с.
10. Liu T., Moore A.W., Gray A. New Algorithms for Efficient High Dimensional Non-Parametric Classification // Journal of Machine Learning Research, 2006, pp. 1135-1158.
11. Omohundro S.N. Five Ball Tree Construction Algorithms 1989, Technical Report.
12. Sproull R.F. Refinements to Nearest Neighbor Searching. Technical Report, International Computer Science, ACM (18) 9, pp. 507-517.
13. Li S.Z., Chan K.L. Performance Evaluation of The NFL Method in Image Classification and Retrieval // IEEE Trans On Pattern Analysis and Machine Intelligence, Vol. 22, 2000.
14. Dhanabal S., Chandramathi S. A Review of various k-Nearest Neighbor Query Processing Techniques // International Journal of Computer Applications, 2011. V. 31, № 7. P. 14-23.