Министерство образоваия и науки Российской Федерации

Федеральное государственное автономное образовательное учреждение
высшего профессионального образования
«Московский физико-технический институт
(государственный университет)»

Факультет управления и прикладной математики

Кафедра проблем передачи информации и анализа данных

ИССЛЕДОВАНИЕ СЛУЧАЙНЫХ ГЕОМЕТРИЧЕСКИХ ПОЛЕЙ ПОСРЕДСТВОМ ГРАФОВЫХ МЕТРИК

Выпускная квалификационная работа
(бакалаврская работа)

Направление подготовки: 03.03.01 Прикладные математика и физика

Выполнил:
студент 177 группы _____ Виденеева Анастасия Сергеевна

Научный руководитель:
д. ф.-м. н., главный научный сотрудник _____ Чеботарев Павел Юрьевич

Москва 2015

Оглавление

Введение

Во многих задачах машинного обучения графы используются для моделирования связей между объектами. Например, анализ социальных графов и сетей, создание рекомендательных систем, транспортные задачи.

Наиболее важная часть решения подобных задач — это выбор способа измерения расстояния между вершинами. Для этого используются различные метрики, которые отражают разные свойства графа. Наиболее простой способ определить расстояние — кратчайший путь — не всегда дает хорошие результаты, потому что этот метод не учитывает связи, которые длиннее, чем самая короткая, и их количество. Другая распространенная метрика — resistance distance, как и пропорциональная ей commute time distance, учитывает все возможные пути между вершинами. Однако, в работе [1] было показано, что при росте количества вершин в графе данные метрики сходятся к функциям, зависящим от степеней вершин и не отражающим глобальных свойств графа. Были предложены другие способы измерить расстояние между вершинами, большинство из которых представляет собой параметрические семейства и при предельных значениях параметров сходится либо к расстоянию кратчайшего пути, либо к resistance distance. В данной работе изучается поведение этих метрических семейств.

Целью работы является исследование близости метрик к исходному евклидовому расстоянию между вершинами графа для четырех типов случайных геометрических графов: ε-графов, симметричных графов ближайших соседей, взаимных графов ближайших соседей и полных графов с гауссовским распределением весов ребер, в зависимости от параметра метрики.

Для этого разрабатывается модель, позволяющая генерировать графы и вычислять расстояния между их вершинами с помощью различных метрик и критерии сравнения метрик с евклидовым расстоянием. Затем проводятся эксперименты, в ходе которых исследуется зависимость поведения метрик от типа графа и его параметров и вычисляются оптимальные в смысле выбранных критерев качества значения параметра метрик для каждого типа графов.

Глава 1

Постановка задачи

1.1 Основные определения

Пусть $G = (V, E)$ — неориентированный граф с множеством вершин V и множеством ребер E, n - число вершин. Матрицу смежности невзвешенного графа будем обозначать $A = (a_{ij})$, где $a_{ij} = 1$, если ребро $(v_i, v_j) \in E$ и $a_{ij} = 0$ в противном случае. Для взвешенных графов будем хранить в этой матрице веса ребер: $a_{ij} = w(v_i, v_j)$. Обозначим D матрицу степеней вершин графа G.

Также в работе используются понятия спектрального радиуса матрицы: $\rho(A) = \max_i \lambda_i(A)$ и лапласиана графа: $L = D - A$

Определение 1 *Метрикой на множестве* X называется функция $d : X^2 \to \mathbb{R}$ такая, что для любых x, y, $z \in X$ выполнены следующие утверждения:

1. $d(x,y) = 0$ тогда и только тогда, когда $x = y$
2. $d(x,y) + d(x,z) - d(y,z) \geq 0$ (неравенство треугольника)

Из этого определения следует, что для любых x, $y \in X$:

1. $d(x,y) = d(y,x)$ (симметричность)
2. $d(x,y) \geq 0$ (неотрицательность)

На практике графовые метрики часто получают из функций близости. Они широко применяются в теории графов и сетей, исследовании марковских процессов и анализе статистических моделей. В данной работе рассматриваются два класса функций близости: Σ-*близости* и *передаточные меры*. Приведем определения этих классов и ряд теорем, показывающих связь между ними и метриками.

Определение 2 Пусть X — непустое множество и $\Sigma \in \mathbb{R}$. Функция $\sigma : X^2 \to \mathbb{R}$ называется Σ-*близостью* на A, если для любых x, y, $z \in X$ выполняются следующие условия:

1. $\sum\limits_{t \in X} \sigma(x,t) = \Sigma$
2. $\sigma(x,y) + \sigma(x,z) - \sigma(y,z) \leq \sigma(x,x)$, где при $z = y$ и $x \neq y$ неравенство строгое.

В работе [2] было доказано, что между метриками и Σ-proximities на множестве X существует взаимно однозначное соответствие.

Определение 3 Пусть G - мультиграф с набором вершин V. Функция $d : V * V \to \mathbb{R}$ называется *граф-геодезической (graph-geodetic), или разрезно-аддитивной (cutpoint addictive)*, если $d(i,j) + d(j,k) = d(i,k)$ выполнено тогда и только тогда, когда в графе G путь, соединяющий вершины i и k, проходит через вершину j.

Определение 4 Говорят, что матрица $S = (s_{ij}) \in \mathbb{R}^{n \times n}$ задает *передаточную меру* $s(i,j) = s_{ij}$ на вершинах $i,j \in V$ графа G, если ее элементы удовлетворяют передаточному неравенству

$$s_{ij} s_{jk} \leq s_{ik} s_{jj}.$$

Это неравенство является аналогом неравенства треугольника для мер близости.

Теорема Пусть $S = (s_{ij}) \in \mathbb{R}^{n \times n}$ задает транзитивную меру на графе G и все недиагональные элементы этой матрицы положительны. Тогда матрица $D = (d_{ij})_{n \times n}$, определенная как

$$D \ = \ (h\mathbf{1}^\intercal + \mathbf{1}h^\intercal - H - H^\intercal)/2,$$

где H получается поэлементным логарифмированием матрицы S, является матрицей расстояний на $V(G)$. Более того, это расстояние будет cutpoint addictive.

Доказательство этой теоремы можно найти в [3].

В данной работе расстояние между вершинами в графе задается матрицей расстояний $D = (d_{ij})$, которую получают из определенным образом заданных мер близости $H = (h_{ij})$ с помощью преобразования

$$D \ = \ (h\mathbf{1}^\intercal + \mathbf{1}h^\intercal - H - H^\intercal)/2,$$

где h — вектор-диагональ матрицы H.

В некоторых случаях вместо матрицы H можно использовать матрицу H_0, состоящую из логарифмов элементов матрицы H.

1.2 Задача

Пусть G — случайный геометрический граф. В данной работе рассматриваются четыре класса графов: ε-графы, два типа графов ближайших соседей, графы с гауссовским распределением весов ребер. Требуется исследовать близость параметрических семейств графовых метрик на этом графе к евклидовому расстоянию между вершинами графа и найти оптимальные параметры метрик, при которых метрики наилучшим образом приближают это расстояние. Для этого необходимо выбрать критерий сравнения метрик с евклидовым расстоянием.

Также требуется сравнить поведение логарифмических и нелогарифмических метрик.

Проверяется гипотеза о том, что если перед сравнением возвести все элементы матрицы D в некоторую степень из интервала (0,1), то качество приближения евклидового расстояния может улучшиться. Для каждой метрики требуется найти такую степень.

1.3 Исследуемые метрики

В данной работе рассматриваются следующие параметрические семейства графовых метрик:

1. **Маршрутное расстояние (Walk distance)**

   Это параметрическое семейство строится с использованием меры близости

   $$H = (I - tA)^{-1}, \tag{1.1}$$

   где параметр $0 < t < \rho^{-1}$, ρ — спектральный радиус матрицы A. При предельных значениях параметра метрика сходится к shortest path distance и long walk distance. Данное семейство задает Σ-близость, доказательство этого факта в работе [4]. Интерпретацию метрики можно найти в [4]

2. **Логарифмическое маршрутное расстояние (Logarithmic walk distance)**

   Мера H_0 получается поэлементным логарифмированием матрицы H, определяющей Walk distance. Эта матрица задает передаточную меру, доказательство можно найти в работе [5].

3. **e-walk distance**

   Является модификацией Walk distance для взвешенных графов

   Веса ребер рассчитываются по следующей формуле:

   $$w_{ij} = \frac{a_{ij}}{\rho} e^{-\frac{1}{\alpha a_{ij}}}, \tag{1.2}$$

   где a_{ij} - элемент матрицы смежности A, ρ - спектральный радиус A, $\alpha > 0$ - параметр метрики.

   Свойства данного семейства и доказательство того, что оно является Σ-близостью, можно найти в работе [4].

4. **Forest distance**

   Корневое дерево (rooted tree) — связный ациклический граф, одна вершина в котором отмечена как корень. *Корневой лес (rooted forest)* — граф, все связные компоненты которого являются rooted trees.

   Рассмотрим взвешенный граф G. Обозначим за $w(G)$ произведение весов его ребер. Для графа без ребер $w(G) = 1$. Если S — набор графов, то $w(S) = \sum\limits_{G \in S} w(G)$. В случае, когда S — пустое множество, $w(S) = 0$. Если множество S состоит из невзвешенных графов, то $w(S) = |S|$.

   Введем следующие обозначения:

1. $F = F(G)$ - множество остовных корневых лесов (spanning rooted forests) графа G;
2. $F_{i,j} = F_{i,j}(G)$- множество таких остовных корневых лесов, что вершина i принадлежит дереву с корнем j;
3. $F_{i,j}^{(p)} = F_{i,j}^{(p)}(G)$ - подмножество таких остовных корневых лесов множества $F_{i,j}$, которые содержат ровно p ребер.

Пусть

$$f = w(F), \;\; f_{i,j} = w(F_{i,j}), \;\; f_{i,j}^{(p)} = w(F_{i,j}^{(p)}),$$

где $i,j \in V(G)$ и $0 \leq p < n$.

Теперь рассмотрим матрицу $Q = (I + L)^{-1}$.

Согласно *Matrix forest theorem*, такая матрица существует для любого взвешенного мультиграфа и ее элементы равны $q_{i,j} = f_{i,j}/f, \;\; i, j = 1, 2 \ldots n$. Матрицу Q можно рассматривать как меру близости.

Добавим зависимость от параметра:

$$H = (I + tL)^{-1}, \tag{1.3}$$

где параметр $t > 0$, а L — лапласиан графа.

При $t \to \infty$ данная метрика сходится к resistance distance. Данное семейство задает Σ-близость и описано в [6].

5. **Logarithmic forest distance**

H получена поэлементным логарифмированием матрицы близости для forest distance. Эта матрица задает транзитивную меру, доказательство этого факта и свойства метрики можно найти в работах [5], [7] и [6].

6. **Communicability distance**

Communicability между вершинами p и q в графе G - это взвешенная сумма всех блужданий, которые начинаются в p и заканчиваются в q, при этом чем короче блуждание, тем больше его вес. Если A - матрица смежности графа, то Communicability между вершинами p и q - это соответствующий элемент матрицы e^A.

Данное определение имеет простую физическую интерпретацию. Рассмотрим граф как систему из шариков массой m, соединенных пружинами с константой $m\omega^2$. Затем вся эта система погружается в жидкость с температурой T. Под воздействием температуры шарики начинают осциллировать.

Гамильтониан системы имеет следующий вид:

$H = \sum\limits_i \left(\frac{p_i^2}{2m} + (K - k_i)\frac{m\omega^2 x_i^2}{2} \right) + \frac{m\omega^2}{2} \sum\limits_{i,j:i<j} A_{ij}(x_i - x_j)^2,$

где k_i - степень вершины i, $K \geq \max_i k_i$, x_i - координата i-го шарика, характеризующая его отклонение от положения равновесия $x_i = 0$. Тогда в предположении, что система подчиняется законам квантовой механики, элемент G_{pq} - это термальная функция Грина осциллирующей системы когда обратная температура равна нулю. Следовательно, G_{pp} показывает, какая часть возбуждения узла p передается в систему до того, как оно возвращается обратно и угасает, а элемент G_{pq} показывает, какая часть этого возбуждения передается от вершины p к вершине q.

Функция близости, соответствующая данному расстоянию имеет вид:

$$H = e^{tA}, \tag{1.4}$$

параметр $t > 0$

Данное семейство задает Σ-близость. Его свойства описаны в работе [8].

7. **Logarithmic communicability distance**

   H получена поэлементным логарифмированием матрицы близости для communicability distance. Данное семейство задает транзитивную меру.

8. **Free energy distance**

   Это семейство метрик, зависящее от параметра β, было рассмотрено в работе [9]. Физический смысл параметра - температура. Данное расстояние вычисляется следующим образом:

   $P^{ref} = D^{-1}A$, $D = \mathrm{diag}(Ae)$, то есть P^{ref} - матрица commute time расстояний между вершинами графа

   $W = P^{ref} \circ e^{-\beta C}$, где $\circ$ означает поэлементное умножение, а элементы матрицы C $c_{ij} = 1/a_{ij}$

   $Z = (I - W)^{-1}$

   $Z^h = ZD_h^{-1}$, $D_h = \mathrm{diag}(Z)$

   $\Phi = -\frac{1}{\beta}\log Z^h$ - матрица свободных энергий, логарифмирование поэлементное

$$D^{FE} = (\Phi + \Phi^T)/2 \tag{1.5}$$

   Данное расстояние стремится к расстоянию кратчайшего пути при $\beta \to \infty$ и к commute time при $\beta \to 0^+$.

9. **Shortest path distance**

   Кратчайшим путем между двумя вершинами графа называют такой путь между этими вершинами, что сумма длин ребер (величин, обратных весам), из которых он состоит, минимальна.

Существует несколько способов вычисления кратчайшего пути, в данной работе используется алгоритм Флойда - Уоршелла [10].

10. **Resistance distance**

Резисторное расстояние между двумя вершинами эквивалентно эффективному сопротивлению между соответствующими точками в электрической цепи, полученной из графа G заменой ребер на резисторы, сопротивление которых совпадает с весом ребер.

$$H = (L + J)^{-1}, \tag{1.6}$$

где L - лапласовская матрица, J - матрица, все элементы которой равны $\frac{1}{n}$, гдк n - число вершин. Данное семейство задает Σ-близость.

11. **Avrachenkov distance**

Данное семейство мер близости было предложено в [11]. Оно возникло при исследовании способов решения задачи классификации с частичным привлечением учителя (semi-supervised classification), которые основаны на использовании графов. В данной работе оно впервые рассматривается как функция близости.

$$H = (1 - a)(I - aD^{-\sigma}AD^{\sigma-1})^{-1}, \tag{1.7}$$

где $a = \frac{2}{2+\mu}$, μ - параметр регуляризации, который позволяет регулировать баланс между точностью классификации и гладкостью классифицирующей функции. Параметр σ позволяет использовать общую формулу для трех методов классификации с частичным привлечением учителя. При $\sigma = 1$ получаем метод, основанный на использовании стандартного лапласиана графа, $\sigma = 0.5$ - нормированного лапласиана, случай $\sigma = 0$ соответствует PageRank.

D - матрица степеней вершин. В случае взвешенных графов вычисляется как сумма весов ребер, инцидентных данной вершине.

12. **Logarithmic Avrachenkov distance**

Данная мера близости вычисляется с помощью поэлементного логарифмирования элементов матрицы H для метрики Авраченкова.

Глава 2

Исследование поведения метрик для различных типов графов

2.1 Генерация вершин графов

В данной работе вершины графа генерировались с помощью смеси гауссовских распределений. Основной случай: четыре двумерные гауссианы, центры которых расположены симметрично относительно начала координат, дисперсии и количество точек равны.

2.2 Генерация невзвешенных графов

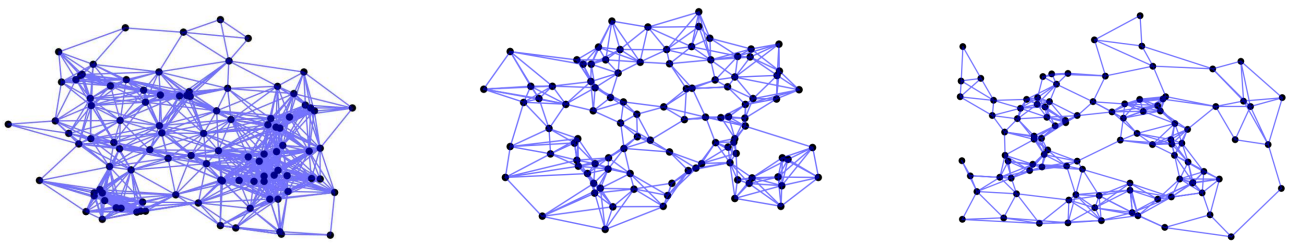
В данной работе рассматривались три класса случайных геометрических невзвешенных графов графов:

1. **ε-графы**: вершины соединяются ребром в том случае, когда евклидово расстояние между ними не превышает заданного параметра ε.

2. **Симметричные графы ближайших соседей**: между двумя вершинами проводится ребро в том случае, если хотя бы одна из них попадает в множество k ближайших соседей другой; параметр k задан.

3. **Графы взаимных ближайших соседей**: две вершины соединяются ребром, если обе они попадают в множество k ближайших соседей друг друга; параметр k задан.

Параметр графа (ε или k) выбирался таким образом, чтобы граф оказался связным с высокой вероятностью. Это делалось потому, что наибольший интерес для машинного обучения представляют именно связные графы.



а) б) в)

Рисунок 2.1: Примеры невзвешенных графов со 100 вершинами.
Слева направо: ε-граф, симметричный граф ближайших соседей ($k = 6$), граф взаимных ближайших соседей ($k = 9$).

2.3 Генерация взвешенных графов

В данной работе взвешенные графы представлены гауссовскими графами. Это полные графы, в которых вес ребра между вершинами i и j определяется по формуле $w_{ij} = \exp(-||v_i - v_j||^2/\sigma^2)$, где параметр $\sigma > 0$ задан.

2.4 Сравнение метрик

Для каждого типа графов для различных значений параметра метрик вычисляются матрицы расстояний для каждой метрики, описанной в главе 1. Чтобы оценить «качество» метрик, они сравниваются с евклидовым расстоянием между вершинами графа. Для этого из элементов матрицы расстояний метрики $D_{metrics}$ и матрицы евклидовых расстояний D_{euclid} составляются векторы d_m и d_e, которые сравниваются между собой следующими способами:

– Коэффициент корреляции Пирсона

– Коэффициент ранговой корреляции Спирмена

– Векторная 1-норма для вектора $d_m^{norm} - d_e^{norm}$

– Векторная 2-норма для вектора $d_m^{norm} - d_e^{norm}$

где индекс $norm$ означает, что вектор с помощью линейного преобразования приведен к нулевому среднему и единичной дисперсии.

Глава 3

Результаты

3.1 Незвешенные графы

3.1.1 ε-графы

Результаты экспериментов представлены на графиках. Использовались графы на 250 вершинах. Во всех случаях по оси x отложены значения параметра семейства. Для удобства все параметры были отнормированы на отрезок [0,1] с помощью дробно-линейного преобразования. В случае коэффициентов корреляции правый рисунок показывает увеличенную область больших значений коэффициента (> 0.8).

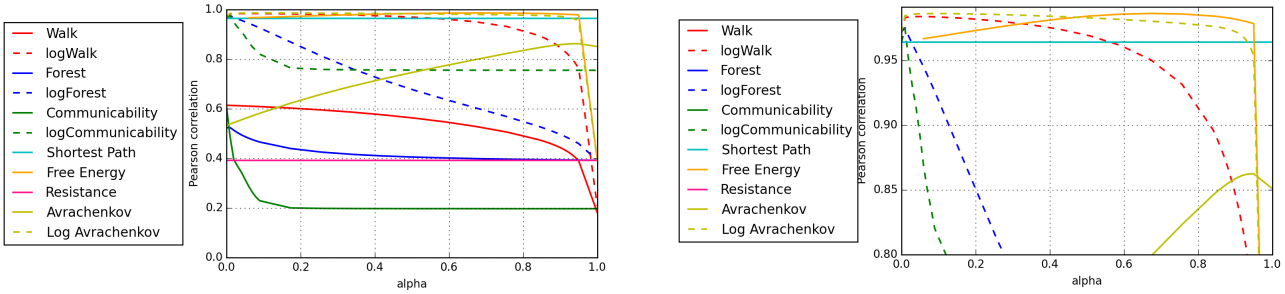


Рисунок 3.1: Корреляции Пирсона для ε-графов

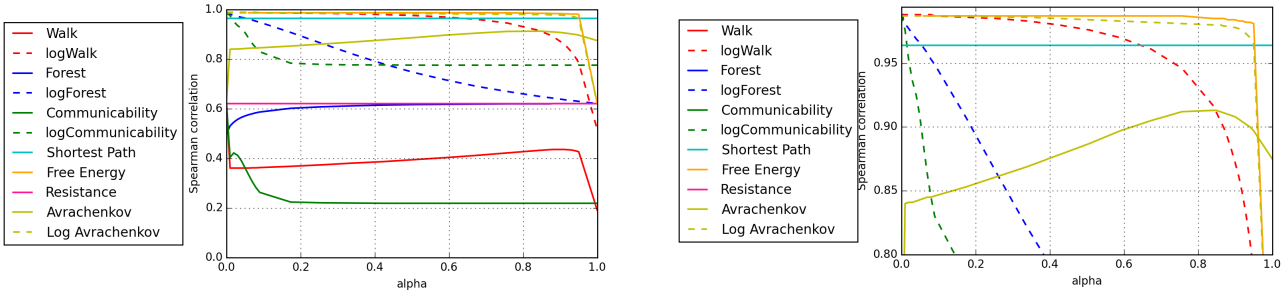


Рисунок 3.2: Корреляции Спирмена для ε-графов

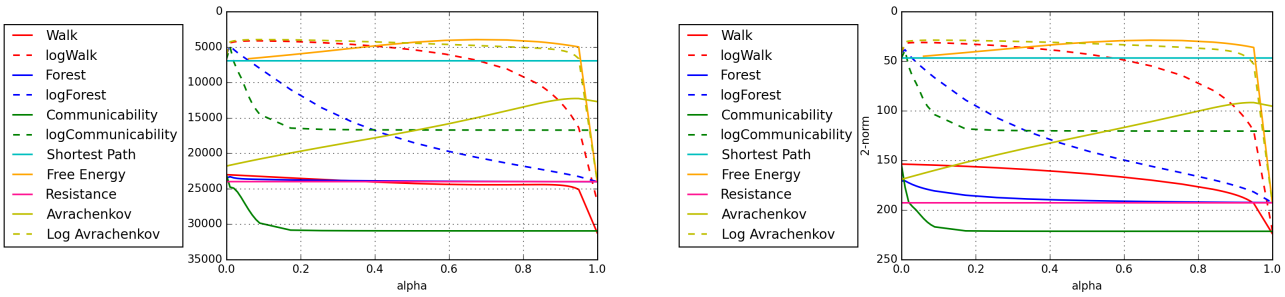


Рисунок 3.3: Матричные нормы для ε-графов: слева - 1-норма, справа - 2-норма

Значения параметров, при которых метрики лучше всего приближают евклидово расстояние, и значение коэффициента корреляции Пирсона для данных параметров, приведены в таблице:

Таблица 3.1: Параметры метрик для ε-графов

Метрика	Значение параметра из [0,1]	Корреляция Пирсона
Walk	0.35	0.613
Log Walk	0.15	0.984
Forest	1.0	0.525
Log Forest	0.005	0.975
Communicability	0.02	0.613
Log Communicability	0.01	0.975
Shortest Path	не зависит	0.964
Resistance	не зависит	0.392
Free Energy	0.7	0.986
Avrachenkov	0.87	0.863
Log Avrachenkov	0.08	0.986

Результаты вычисления максимальных корреляций при возведении матрицы D в степени p, отличные от 1.0 (в таблице показаны только те метрики, которые позволяли получить хорошее приближение евклидового расстояния в предыдущем эксперименте):

Таблица 3.2: Зависимость максимальной корреляции от степени для ε-графов

Метрика	$p = 0.25$	$p = 0.5$	$p = 0.75$	$p = 1.0$
Log Walk	0.975	0.979	0.982	0.984
Log Forest	0.971	0.973	0.974	0.975
Log Communicability	0.974	0.974	0.975	0.975
Shortest Path	0.955	0.960	0.963	0.964
Free Energy	0.982	0.984	0.985	0.986
Avrachenkov	0.862	0.862	0.863	0.863
Log Avrachenkov	0.981	0.984	0.985	0.986

3.1.2 Симметричные графы ближайших соседей

Результаты экспериментов представлены на графиках. Использовались графы на 250 вершинах, параметр $k = 8$. Во всех случаях по оси x отложены значения параметра семейства. Для удобства все параметры были отнормированы на отрезок [0,1] с помощью дробно-линейного преобразования. В случае коэффициентов корреляции правый рисунок показывает увеличенную область больших значений коэффициента (> 0.8).

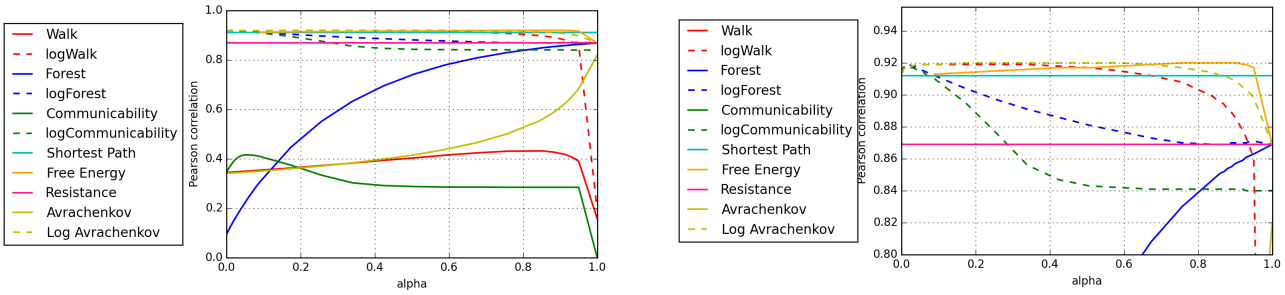


Рисунок 3.4: Корреляции Пирсона для симметричных графов ближайших соседей

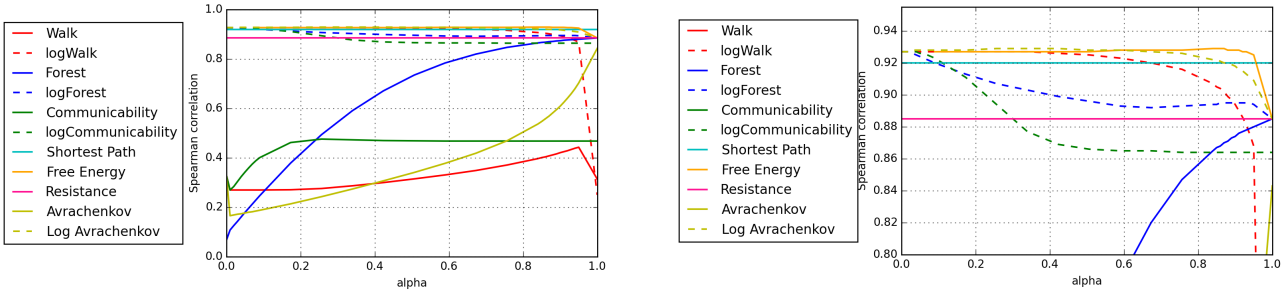


Рисунок 3.5: Корреляции Спирмена для симметричных графов ближайших соседей

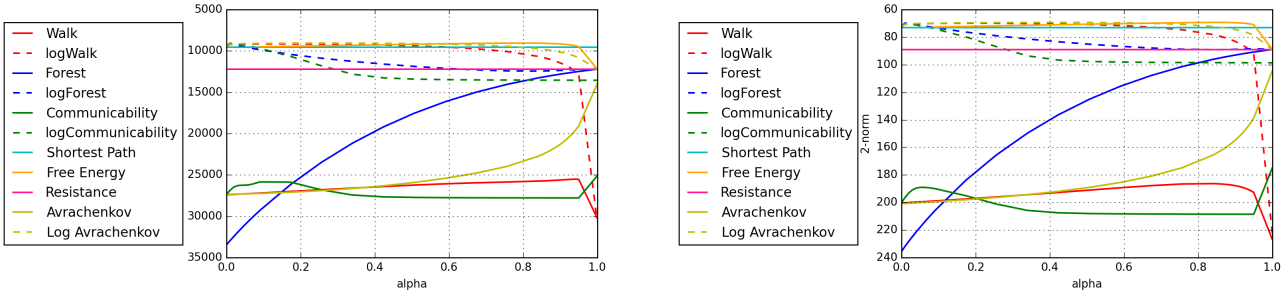


Рисунок 3.6: Матричные нормы для симметричных графов ближайших соседей: слева - 1-норма, справа - 2-норма

Значения параметров, при которых метрики лучше всего приближают евклидово расстояние, и значение коэффициента корреляции Пирсона для данных параметров, приведены в таблице:

Таблица 3.3: Параметры метрик для симметричных графов ближайших соседей

Метрика	Значение параметра из [0,1]	Корреляция Пирсона
Walk	0.87	0.432
Log Walk	0.18	0.919
Forest	1.0	0.869
Log Forest	0.005	0.919
Communicability	0.3	0.416
Log Communicability	0.01	0.918
Shortest Path	не зависит	0.912
Resistance	не зависит	0.869
Free Energy	0.85	0.920
Avrachenkov	0.95	0.813
Log Avrachenkov	0.051	0.920

При возведении элементов матрицы D в степени, отличные от 1.0, качественное поведение корреляций такое же, как в случае ε-графов.

3.1.3 Графы взаимных ближайших соседей

Результаты экспериментов представлены на графиках. Использовались графы на 250 вершинах, параметр $k = 12$. Во всех случаях по оси x отложены значения параметра семейства. Для удобства все параметры были отнормированы на отрезок [0,1] с помощью дробно-линейного преобразования. В случае коэффициентов корреляции правый рисунок показывает увеличенную область больших значений коэффициента (> 0.8).

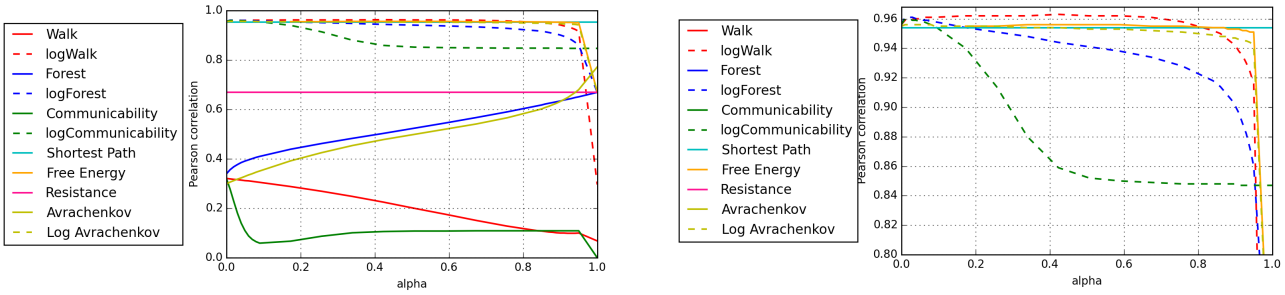


Рисунок 3.7: Корреляции Пирсона для графов взаимных ближайших соседей

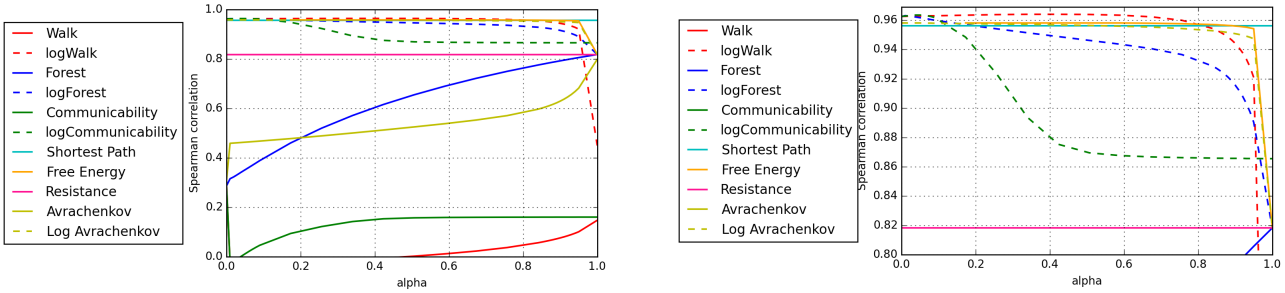


Рисунок 3.8: Корреляции Спирмена для графов взаимных ближайших соседей

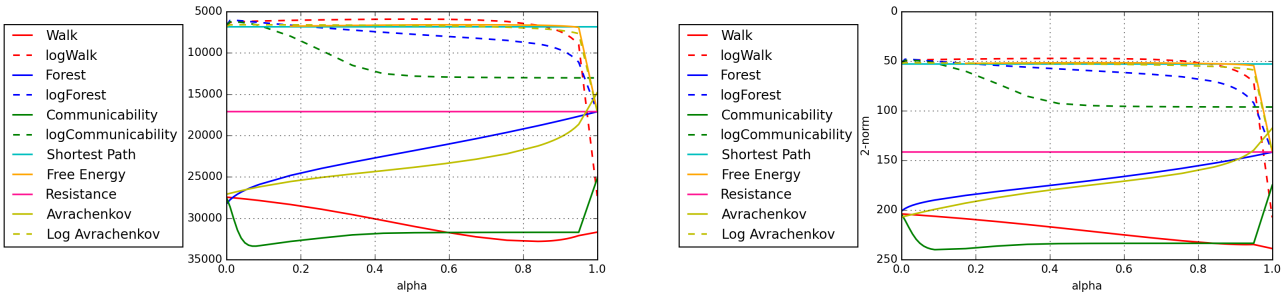


Рисунок 3.9: Матричные нормы для графов взаимных ближайших соседей: слева - 1-норма, справа - 2-норма

Значения параметров, при которых метрики лучше всего приближают евклидово расстояние, и значение коэффициента корреляции Пирсона для данных параметров, приведены в таблице:

Таблица 3.4: Параметры метрик для графов взаимных ближайших соседей

Метрика	Значение параметра из [0,1]	Корреляция Пирсона
Walk	0.01	0.319
Log Walk	0.38	0.963
Forest	1.0	0.669
Log Forest	0.015	0.961
Communicability	0.9	0.321
Log Communicability	0.025	0.960
Shortest Path	не зависит	0.954
Resistance	не зависит	0.669
Free Energy	0.58	0.956
Avrachenkov	0.95	0.680
Log Avrachenkov	0.035	0.956

При возведении элементов матрицы D в степени, отличные от 1.0, качественное поведение корреляций такое же, как в случае ε-графов.

3.2 Взвешенные графы

Результаты экспериментов представлены на графиках. Использовались графы на 250 вершинах, параметр $\sigma = 5$. Во всех случаях по оси x отложены значения параметра семейства. Для удобства все параметры были отнормированы на отрезок [0,1] с помощью дробно-линейного преобразования. В случае коэффициентов корреляции правый рисунок показывает увеличенную область больших значений коэффициента (> 0.8).

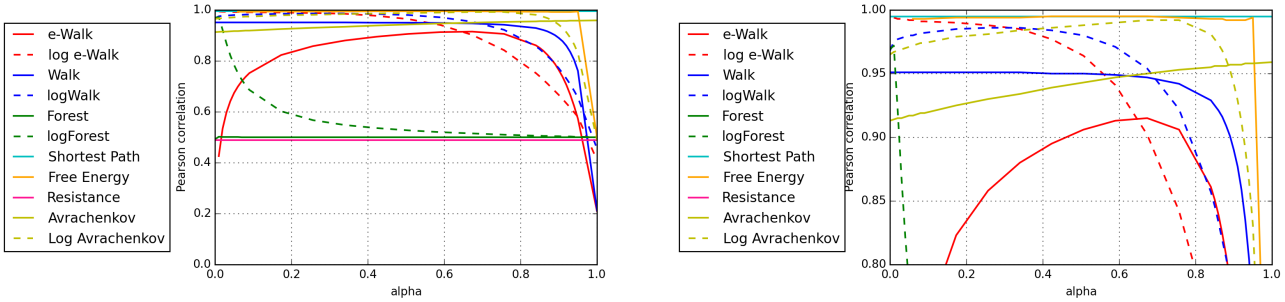


Рисунок 3.10: Корреляции Пирсона для гауссовских графов

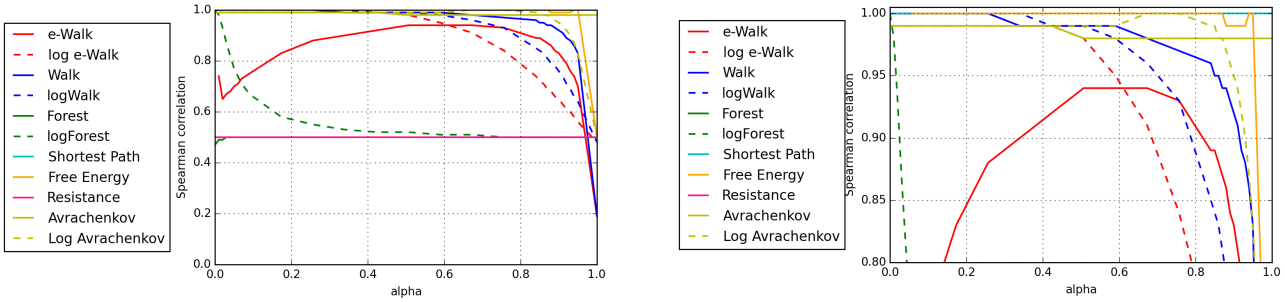


Рисунок 3.11: Корреляции Спирмена для гауссовских графов

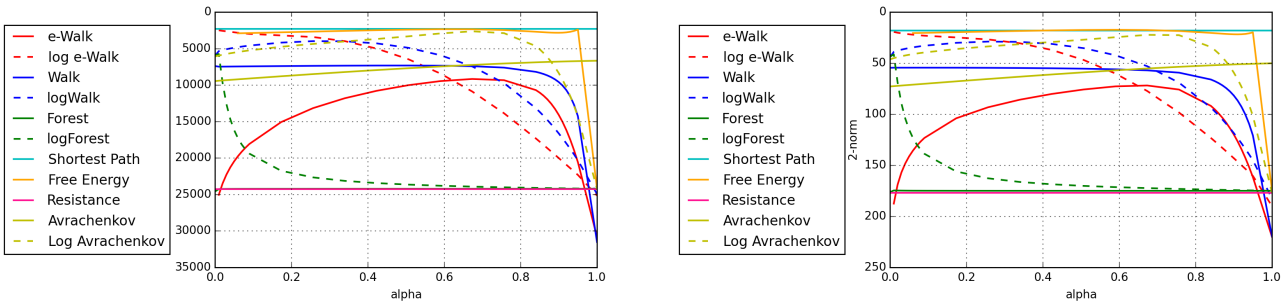


Рисунок 3.12: Матричные нормы для гауссовских графов: слева - 1-норма, справа - 2-норма

Значения параметров, при которых метрики лучше всего приближают евклидово расстояние, и значение коэффициента корреляции Пирсона для данных параметров, приведены в таблице:

Таблица 3.5: Параметры метрик для гауссовских графов

Метрика	Значение параметра из [0,1]	Корреляция Пирсона
Walk	0.87	0.951
Log Walk	0.18	0.986
e-Walk	0.3	0.915
Log e-Walk	0.01	0.994
Forest	1.0	0.501
Log Forest	0.005	0.972
Shortest Path	не зависит	0.995
Resistance	не зависит	0.489
Free Energy	0.85	0.995
Avrachenkov	0.95	0.959
Log Avrachenkov	0.73	0.992

При возведении элементов матрицы D в степени, отличные от 1.0, качественное поведение корреляций такое же, как в случае ε-графов.

3.3 Комментарии к результатам

В экспериментах усреднение производилось по 50 графам с одинаковыми параметрами. Затем усредненные результаты для разных параметров сравнивались между собой.

Максимумы на графиках незначительно меняют свое положение при изменении параметров графа до тех пор, пока он не начинает распадаться на кластеры, после чего значения коэффициентов корреляции в максимумах начинает уменьшаться.

Заметим, что в случае метрики Авраченкова представлены результаты только для $\sigma = 1.0$, для других значений данного параметра зависимость от параметра аналогичная. Это связано с близостью степеней вершин в исследуемых графах.

Для вычисления communicability distance использовалось приближение матричной экспоненты первыми 40 членами ряда.

Выводы

Анализируя результаты экспериментов, можно сделать следующие выводы:

1. Логарифмическое преобразование метрики позволяет значительно улучшить качество приближения евклидового расстояния. Для всех рассмотренных типов графов по всем четырем критериям логарифмические метрики показывают лучшие результаты, чем метрики без логарифма. Данное наблюдение является очень значимым, поскольку в настоящее время логарифмические метрики применяются очень редко.

2. Можно заметить, что качественно поведение графиков зависимости коэффициентов корреляции и векторных норм от параметра метрики схоже. Максимумы на графиках незначительно меняют свое положение при изменении параметров графа (размерность, число гауссиан в смеси, положение их центров и дисперсии, параметр ε или k) до тех пор, пока граф не начинает распадаться на кластеры, после чего значения коэффициентов корреляции в максимумах начинает уменьшаться.

3. Следует отметить метрику Авраченкова. До данной работы эта функция никогда не рассматривалась в качестве метрики, и эксперименты показали, что ее применени на практике имеет смысл: и сама метрика, и ее логарифм позволяют с высокой точностью восстановить евклидовое расстояние между вершинами исходного графа. Заметим, что в данной работе представлены результаты только для $\sigma = 1.0$, потому что для других значений данного параметра зависимость от параметра a аналогичная. Это связано с близостью степеней вершин в исследуемых графах.

4. Несмотря на то, что логарифмические преобразования дают очень хорошие результаты для всех типов графов, в случае гауссовских взвешенных графов расстояние кратчайшего пути позволяет восстановить евклидово расстояние с точностью до константы. Это связано с особенностями определения весов ребер (в случае невзвешенных графов по ребрам можно только понять, меньше ли расстояние между вершинами, чем заданный параметр ε и попадает ли вершина в число k ближайших соседей другой; в то время как для взвешенных графов веса содержат информацию непосредственно о евклидовом расстоянии между вершинами).

5. Гипотеза о том, что возведение элементов матрицы D в степени, отличные от 1.0, может улучшить качество приближения евклидового расстояния, была отвергнута по результатам экспериментов. Для метрик, которые позволяют восстанавливать евклидовое расстояние наилучшим образом, степень 1.0 дает самые лучшие результаты. Метрики, которые при промежуточных значениях параметра имеют невысокие корреляции с евклидовым расстоянием, не представляют интереса для этого исследования, потому что наилучшее приближение евклидового расстояния они дают при предельных значениях параметра, при котором они стремятся к другим метрикам, уже не зависящим от параметра.

6. Можно заметить, что значения параметров, при которых метрики наилучшим образом приближают евклидовое расстояние, похожи для разных типов графов: например, для логарифмической маршрутной метрики корреляции максимальны, а векторные нормы разностей минимальны для параметров из интервала $(0.15, 0.5)$, корреляции с евклидовым расстоянием для метрик logarithmic forest и logarithmic communicability от параметра имеют максимум при небольших значениях параметра, в пределах интервала $(0, 0.03)$, метрика free energy и метрика Авраченкова (не логарифмическая) позволяют наиболее точно восстановить евклидово расстояние при больших значениях параметра.

Заключение

В данной работе было рассмотрено большое число параметрических семейств графовых метрик и исследовано их поведение в зависимости от параметра для четырех типов случайных геометрических графов. Сравнение метрик осуществлялось посредством сравнения каждой из них с евклидовым расстоянием между вершинами графа.

На основании проведенных экспериментов можно сделать вывод, что логарифмическое преобразование позволяет значительно улучшить метрики.

Численные исследования позволили найти значения параметров метрических семейств, которые наиболее интересны для практических приложений, а также было выяснено, что возведение элементов матрицы расстояний в степени, отличные от 1.0, не позволяют получить лучшее приближение евклидового расстояния.

Таким образом, использование графовых метрик, отличных от кратчайшего пути, позволяет приближать евклидовое расстояние между вершинами с высокой точностью и при этом учитывать различные связи (пути) между вершинами, а значит, расстояния, вычисленные с помощью этих метрик, отражают больше информации о структуре графа, чем расстояние кратчайшего пути.

Исходный код для воспроизведения результатов, описанных в данной работе, доступен по адресу https://github.com/vi-nastya/Bachelor-Graph-Metrics

Направления для дальнейших исследований

Данное исследование можно продолжить для других моделей случайных графов. Для случаев, когда евклидово расстояние между вершинами неизвестно, необходимо ввести новый критерий сравнения метрик между собой.

Также интерес представляет применение рассмотренных метрик к задачам классификации и кластеризации на графах.

Список литературы

1. Von Luxburg U., Radl A., Hein M. Hitting and commute times in large random neighborhood graphs // The Journal of Machine Learning Research. — 2014. — Vol. 15, no. 1. — P. 1751–1798.

2. Chebotarev P. Studying new classes of graph metrics // Geometric Science of Information. — Springer, 2013. — P. 207–214.

3. Chebotarev P., Shamis E. On a Duality between Metrics and Σ-Proximities // Automation and Remote Control. — 1998. — Vol. 59, no. 4. — P. 608–612.

4. Chebotarev P. The walk distances in graphs // Discrete Applied Mathematics. — 2012. — Vol. 160, no. 10. — P. 1484–1500.

5. Chebotarev P. The graph bottleneck identity // Advances in Applied Mathematics. — 2011. — Vol. 47, no. 3. — P. 403–413.

6. Chebotarev P. A class of graph-geodetic distances generalizing the shortest-path and the resistance distances // Discrete Applied Mathematics. — 2011. — Vol. 159, no. 5. — P. 295–302.

7. Chebotarev P., Shamis E. The forest metrics for graph vertices // Electronic Notes in Discrete Mathematics. — 2002. — Vol. 11. — P. 98–107.

8. Estrada E. The communicability distance in graphs // Linear Algebra and its Applications. — 2012. — Vol. 436, no. 11. — P. 4317–4328.

9. Kivimäki I., Shimbo M., Saerens M. Developments in the theory of randomized shortest paths with a comparison of graph node distances // Physica A: Statistical Mechanics and its Applications. — 2014. — Vol. 393. — P. 600–616.

10. Floyd R. W. Algorithm 97: shortest path // Communications of the ACM. — 1962. — Vol. 5, no. 6. — P. 345.

11. Generalized optimization framework for graph-based semi-supervised learning / K. Avrachenkov, P. Gonçalves, A. Mishenin, M. Sokol // Proceedings of SIAM Conference on Data Mining (SDM 2012) / SIAM. — Vol. 9. — 2012.