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 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: $a_{ij} = w(v_i, v_j)$.

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-*proximities* и *транзитивные 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-*proximity* 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 *graph-geodetic (bottleneck addictive, cutpoint addictive)*, если $d(i,j) + d(j,k) = d(i,k)$ выполнено тогда и только тогда, когда в графе G любой 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 *транзитивную меру* $s(i,j) = s_{ij}$ на вершинах $i,j \in V$ графа G, если ее элементы удовлетворяют транзитивному 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 разработать 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. **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 *proximity*-меру, доказательство этого факта в работе [5]. Интерпретацию метрики можно найти в [5]

2. **Logarithmic walk distance**

   Мера H_0 получается поэлементным логарифмированием матрицы H, определяющей Walk distance. Эта матрица задает транзитивную меру, доказательство можно найти в работе [4].

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 *proximity*-мерой, можно найти в работе [5].

4. **Forest distance**

   Rooted tree — связный ациклический граф, одна вершина в котором отмечена как корень. *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 $w(S) = \sum_{G \in S} w(G)$. 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 spanning rooted forests графа G;
2. $F_{i,j} = F_{i,j}(G)$- множество таких spanning rooted forests, что вершина i принадлежит дереву с корнем j;
3. $F_{i,j}^{(p)} = F_{i,j}^{(p)}(G)$ - подмножество таких spanning rooted forests 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 *proximity*-меру и описано в [**?**].

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 [4] и [**?**].

6. **Communicability distance**

   Данное семейство и его свойства подробно описаны в [6].

   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:

$H = \Sigma_i(\frac{p_i^2}{2m} + (K - k_i)\frac{m\omega^2 x_i^2}{2}) + \frac{m\omega^2}{2}\Sigma_{i,j:i<j}A_{ij}(x_i - x_j)^2$

See Estrada p2

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

параметр $t > 0$

Данное семейство задает *proximity*-меру.

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 [7]. Физический смысл параметра - температура. Данное расстояние вычисляется следующим образом:

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

   $W = P^{ref} \circ e^{-\beta C}$, где $\circ$ означает поэлементное умножение, а C - матрица кратчайших расстояний между вершинами графа G

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

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

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

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

   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 весов ребер, из которых он состоит, минимальна.

   Существует несколько способов вычисления кратчайшего пути, в данной работе используется алгоритм Флойда - Уоршелла [8].

10. **Resistance distance**

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

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

где L - лапласовская матрица, J - матрица, все элементы которой равны $\frac{1}{n}$, гдк n - число вершин

11. **Avrachenkov distance**

Данное семейство мер близости было предложено в [9]. Оно возникло при исследовании способов решения задачи классификации с частичным привлечением учителя (semi-supervised classification), которые основаны на использовании графов.

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

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 Авранченкова.

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

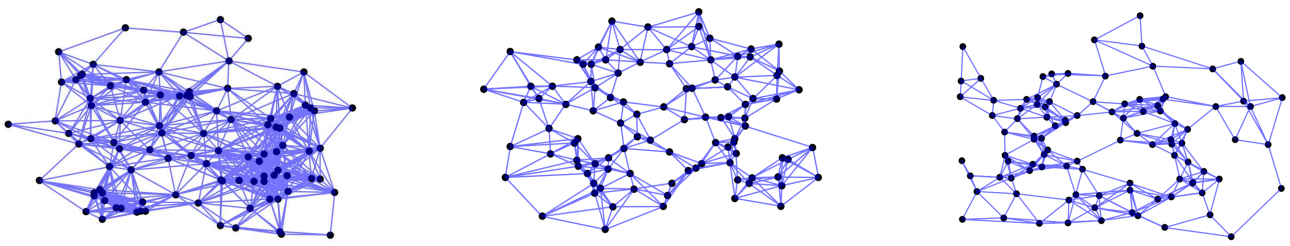
В данной работе рассматривались три класса случайных геометрических невзвешенных графов графов:

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 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).

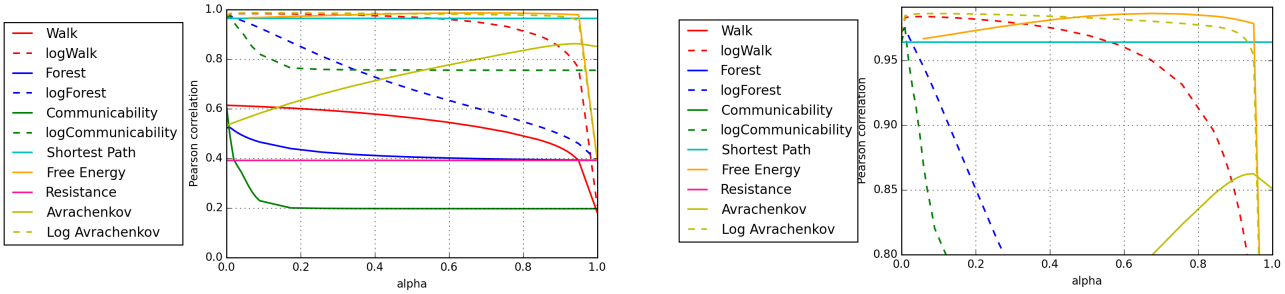


Рисунок 3.1: Корреляции Пирсона для ε-графов

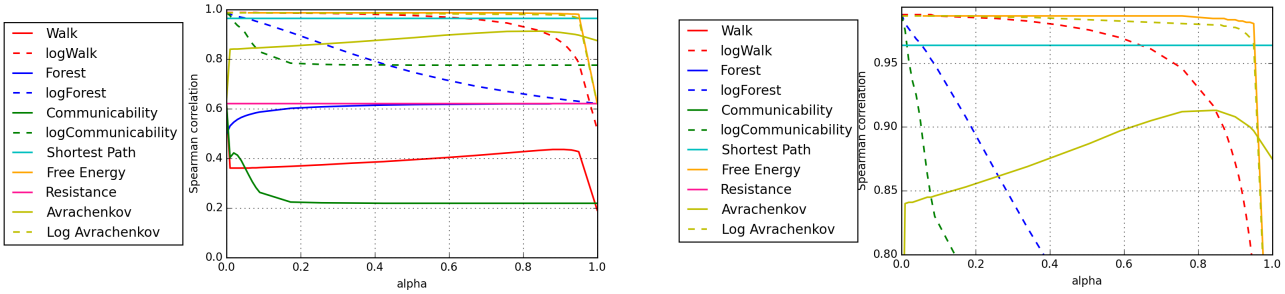


Рисунок 3.2: Корреляции Спирмена для ε-графов

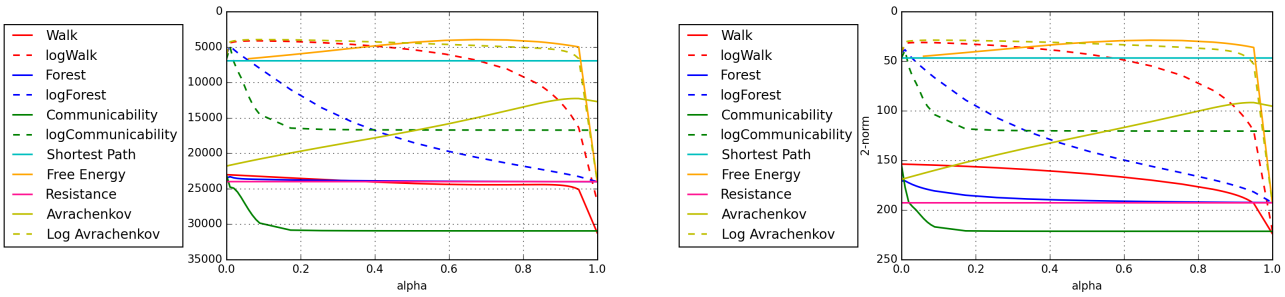


Рисунок 3.3: Матричные нормы для ε-графов: слева - 1-норма, справа - 2-норма

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	0.6
Log Walk	0.15	0.98
Forest	1.0	0.52
Log Forest	0.005	0.98
Communicability	0.02	0.61
Log Communicability	0.01	0.97
Shortest Path	не зависит	0.96
Resistance	не зависит	0.39
Free Energy	0.7	0.97
Avrachenkov	0.87	0.86
Log Avrachenkov	0.08	0.99

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).

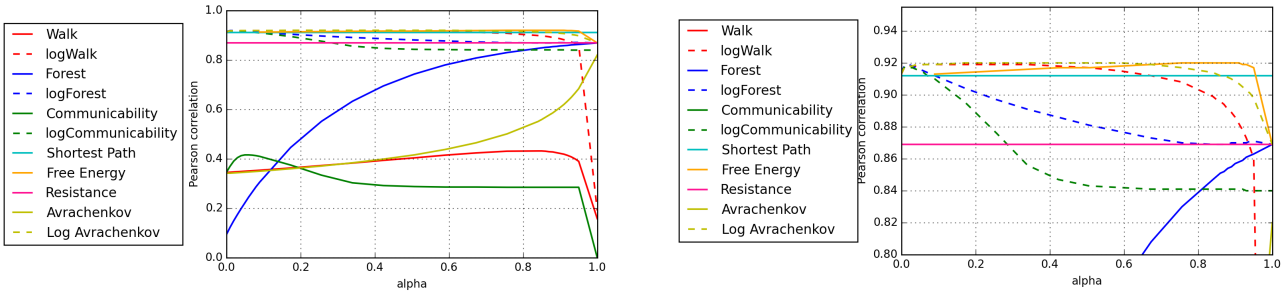


Рисунок 3.4: Корреляции Пирсона для симметричных графов ближайших соседей

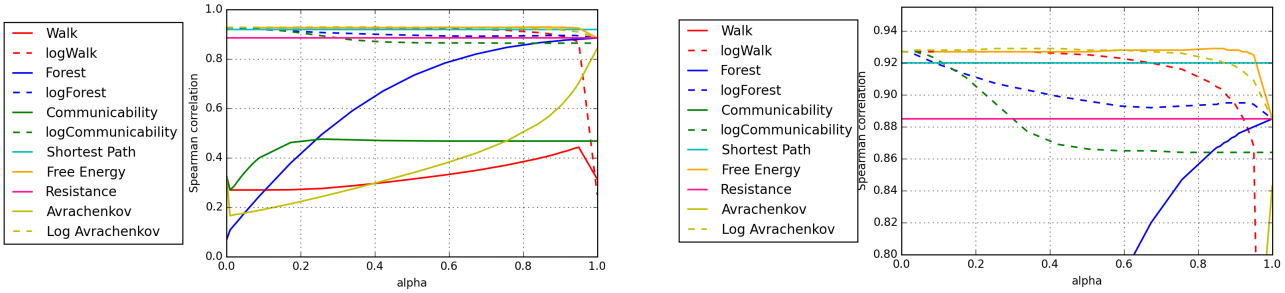


Рисунок 3.5: Корреляции Спирмена для симметричных графов ближайших соседей

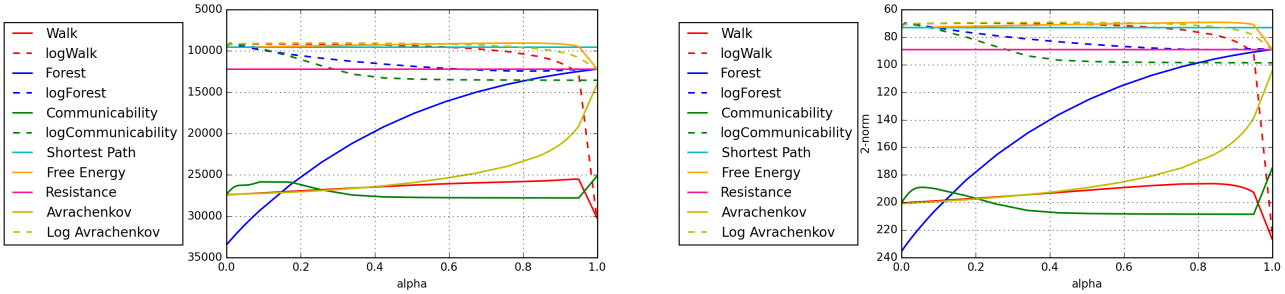


Рисунок 3.6: Матричные нормы для симметричных графов ближайших соседей: слева - 1-норма, справа - 2-норма

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

Таблица 3.2: Параметры метрик для симметричных графов ближайших соседей

Метрика	Значение параметра из [0,1]	Корреляция Пирсона
Walk	0.87	0.43
Log Walk	0.18	0.92
Forest	1.0	0.87
Log Forest	0.005	0.92
Communicability	0.3	0.42
Log Communicability	0.01	0.92
Shortest Path	не зависит	0.91
Resistance	не зависит	0.87
Free Energy	0.85	0.92
Avrachenkov	0.95	0.81
Log Avrachenkov	1.0	0.92

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).

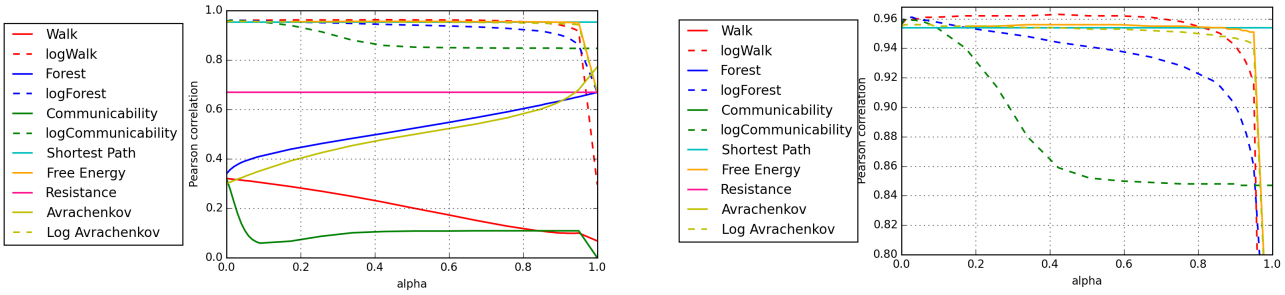


Рисунок 3.7: Корреляции Пирсона для графов взаимных ближайших соседей

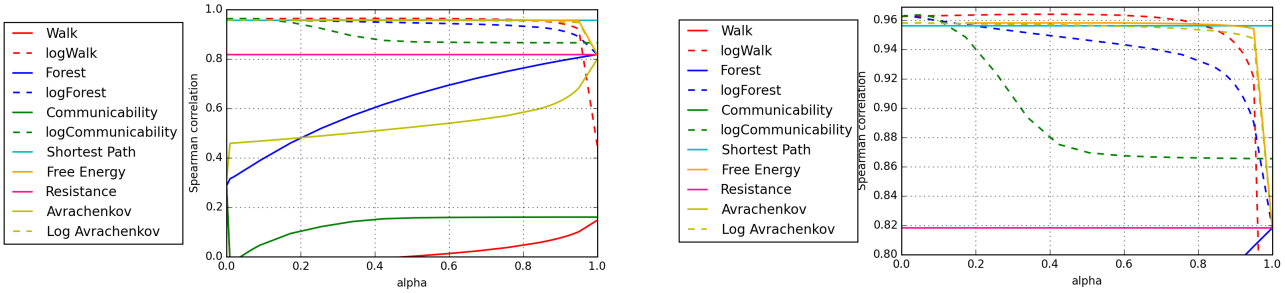


Рисунок 3.8: Корреляции Спирмена для графов взаимных ближайших соседей

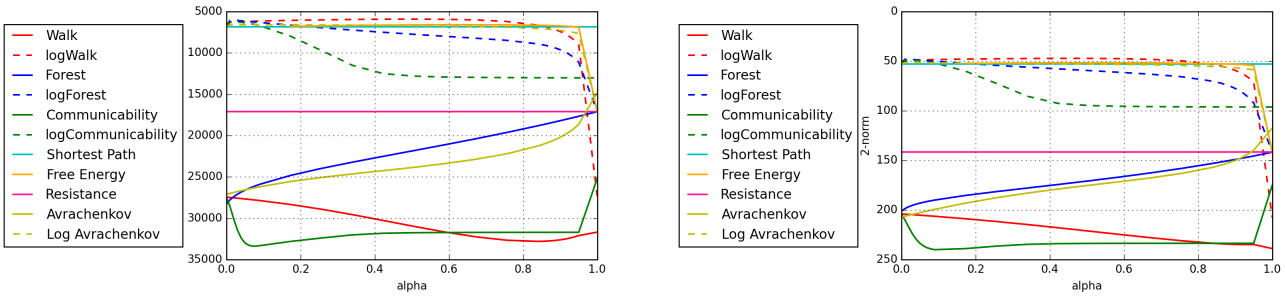


Рисунок 3.9: Матричные нормы для графов взаимных ближайших соседей: слева - 1-норма, справа - 2-норма

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

Таблица 3.3: Параметры метрик для графов взаимных ближайших соседей

Метрика	Значение параметра из [0,1]	Корреляция Пирсона
Walk	0.01	0.32
Log Walk	0.67	0.96
Forest	1.0	0.67
Log Forest	0.015	0.96
Communicability	0.9	0.32
Log Communicability	0.025	0.96
Shortest Path	не зависит	0.95
Resistance	не зависит	0.67
Free Energy	0.58	0.96
Avrachenkov	0.95	0.68
Log Avrachenkov	0.035	0.96

Комментарии к результатам для невзвешенных графов

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.

Также можно заметить, что качественно поведение графиков зависимости коэффициентов корреляции и векторных норм от параметра метрики схоже.

Максимумы на графиках не значительно 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.

Эксперименты показывают, что логарифмические преобразования метрик позволяют лучше приблизить евклидовое расстояние между вершинами графа.

В результате экспериментов для поиска оптимальных степеней выяснилось, что возведение элементов матрицы D в степени, отличные от 1.0, не улучшают качества приближения евклидового расстояния.

Можем видеть, что логарифмические метрики позволяют более точно восстановить евклидово расстояние, чем метрики без логарифма.

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).

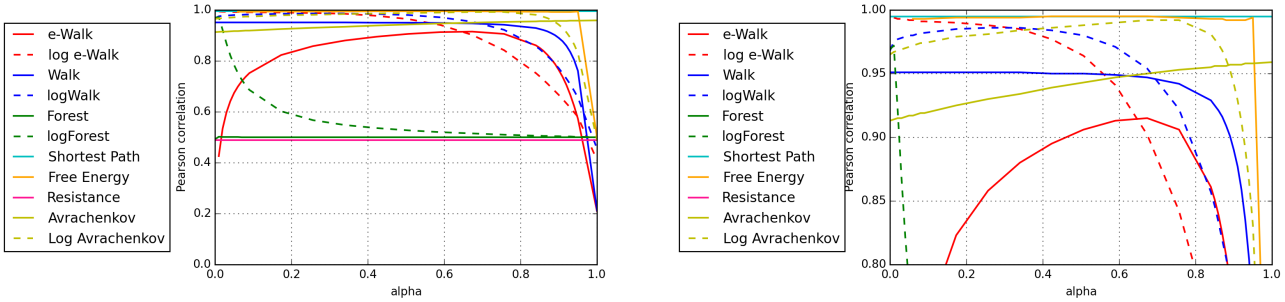


Рисунок 3.10: Корреляции Пирсона для гауссовских графов

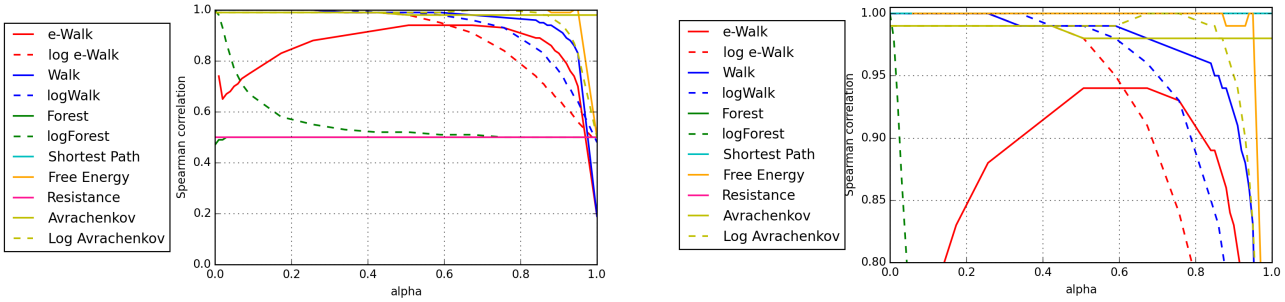


Рисунок 3.11: Корреляции Спирмена для гауссовских графов

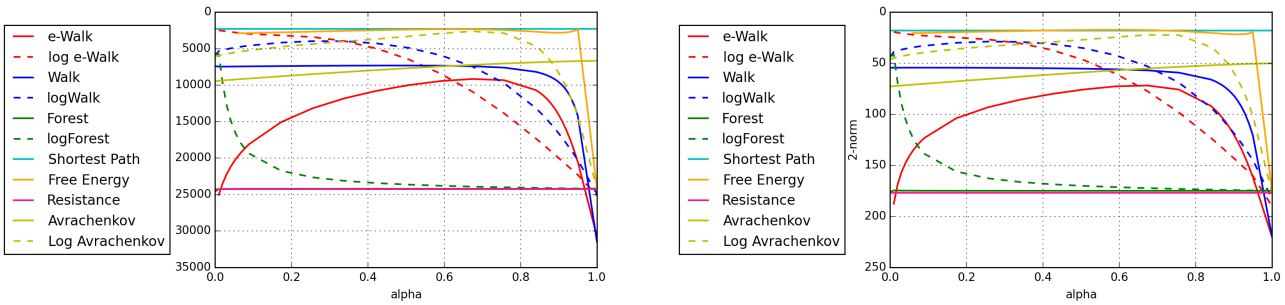


Рисунок 3.12: Матричные нормы для гауссовских графов: слева - 1-норма, справа - 2-норма

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

Таблица 3.4: Параметры метрик для гауссовских графов

Метрика	Значение параметра из [0,1]	Корреляция Пирсона
Walk	0.87	0.95
Log Walk	0.18	0.99
e-Walk	0.3	0.93
Log e-Walk	0.01	0.99
Forest	1.0	0.5
Log Forest	0.005	0.99
Shortest Path	не зависит	0.99
Resistance	не зависит	0.5
Free Energy	0.85	0.99
Avrachenkov	0.95	0.96
Log Avrachenkov	1.0	0.99

Комментарии к результатам для взвешенных графов

Отметим, что, хотя логарифмические преобразования дают очень хорошие результаты, в данном случае расстояние кратчайшего пути позволяет восстановить евклидово расстояние с точностью до константы. Это связано с особенностями определения весов ребер.

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.

Также можно заметить, что, как и в случае невзвешенных графов, качественно поведение графиков зависимости коэффициентов корреляции и векторных норм от параметра метрики схоже.

В результате экспериментов для поиска оптимальных степеней выяснилось, что возведение элементов матрицы D в степени, отличные от 1.0, не улучшают качества приближения евклидового расстояния.

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 метрик. 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 ********

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, Ulrike.* Hitting and commute times in large random neighborhood graphs / Ulrike Von Luxburg, Agnes Radl, Matthias Hein // *The Journal of Machine Learning Research.* — 2014. — Vol. 15, no. 1. — Pp. 1751–1798.

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

3. *Chebotarev, P Yu.* On a Duality between Metrics and Σ-Proximities / P Yu Chebotarev, EV Shamis // *arXiv preprint math/0508183.* — 2005.

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

5. *Chebotarev, Pavel.* The walk distances in graphs / Pavel Chebotarev // *Discrete Applied Mathematics.* — 2012. — Vol. 160, no. 10. — Pp. 1484–1500.

6. *Estrada, Ernesto.* The communicability distance in graphs / Ernesto Estrada // *Linear Algebra and its Applications.* — 2012. — Vol. 436, no. 11. — Pp. 4317–4328.

7. *Kivimäki, Ilkka.* Developments in the theory of randomized shortest paths with a comparison of graph node distances / Ilkka Kivimäki, Masashi Shimbo, Marco Saerens // *Physica A: Statistical Mechanics and its Applications.* — 2014. — Vol. 393. — Pp. 600–616.

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

9. Generalized optimization framework for graph-based semi-supervised learning / Konstantin Avrachenkov, Paulo Gonçalves, Alexey Mishenin, Marina Sokol // Proceedings of SIAM Conference on Data Mining (SDM 2012) / SIAM. — Vol. 9. — 2012.

1. Von Luxburg U., Radl A., Hein M. Hitting and commute times in large random neighborhood graphs //The Journal of Machine Learning Research. – 2014. – Т. 15. – №. 1. – С. 1751-1798.

1. P. Chebotarev. Studying new classes of graph metrics // F. Nielsen and F. Barbaresco (Eds.), Proceedings of the SEE Conference “Geometric Science of Information” (GSI 2013). Lecture Notes in Computer Science, LNCS 8085. Springer, Berlin, 2013. P. 207–214.

2. Chebotarev P. Y., Shamis E. V. On a Duality between Metrics and *Sigma*-Proximities //arXiv preprint math/0508183. – 2005.

2. Hitting and Commute Times in Large Random Neighborhood Graphs, U. von Luxburg, A. Radl, M. Hein. //Journal of Machine Learning Research 15 (2014) 1751-1798

3. A class of graph-geodetic distances generalizing the shortest-path and the resistance distances, P. Chebotarev. //Discrete Applied Mathematics 159 (2011) 295–302

4. Shortest path distance in random k-nearest neighbor graphs, M. Alamgir, U. von Luxburg (2012)

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

6. E. Estrada. The communicability distance in graphs // Linear Algebra and its Applications. 2012. V. 436. P. 4317–4328.

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

8. R.W. Floyd. Algorithm 97: Shortest path. //Communication of the ACM 5(6):345, 1962.

9. K. Avranchenkov, A. Mishenin, P. Gonçalves, M. Sokol. Generalized Optimization Framework for Graph-based Semi-supervised Learning // In: Proceedings of SIAM Conference on Data Mining (SDM 2012)

10. Chebotarev Walk distances in graphs

11. Chebotarev. The forest metric for graph vertices.

12 Chebotarev P. The graph bottleneck identity //Advances in Applied Mathematics. – 2011. – Т. 47. – №. 3. – С. 403-413.