МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»
Кафедра дискретной математики и информационных технологий

АНАЛИЗ ДИНАМИКИ ХАРАКТЕРИСТИК СОЦИАЛЬНЫХ СЕТЕЙ МЕТОДАМИ ТЕОРИИ СЛУЧАЙНЫХ ГРАФОВ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента 2 курса 271 группы
направления 09.04.01 — Информатика и вычислительная техника
факультета КНиИТ
Ионова Кирилла Игоревича

Научный руководитель
к. ф.-м. н., доцент _____ И. Д. Сагаева

Заведующий кафедрой
доцент, к. ф.-м. н. _____ Л. Б. Тяпаев

Саратов 2023

ВВЕДЕНИЕ

Сети являются неотъемлемой частью человеческой жизни. С помощью графов таких сетей можно описать различные сферы: социальные ресурсы, биологические процессы и транспортные схемы. Данные сети постоянно растут и развиваются, с увеличением количества узлов и связей между ними. Возникает необходимость в эмпирическом исследовании, чтобы понять принципы развития таких сетей и анализировать, как они меняются со временем и в зависимости от различных параметров.

Цель данной работы состоит в исследовании локальных и глобальных характеристик сетей, которые описываются при помощи моделей случайных графов. Для достижения этой цели были поставлены следующие задачи:

- Разработка программного кода, позволяющего генерировать графы на основе нескольких моделей случайных графов.
- Анализ изменений локальных и глобальных характеристик в полученных сетях;
- Сравнение глобальных характеристик фрагмента сети Facebook с моделями случайных графов;
- Разработка приложения, которое занимается сбором данных в социальной сети Twitter;
- Сравнение графов, построенных по случайным моделям, с графом, созданным на основе собранных данных.

В данной работе рассматриваются модели случайных графов, такие как Эрдёша-Реньи, Барабаши-Альберта и Холма-Кима. Мы анализируем локальные характеристики, такие как степень вершины, коэффициент кластеризации, средняя степень соседних узлов и индекс дружбы. Среди глобальных характеристик рассматриваем среднюю степень вершин, средний коэффициент кластеризации сети, глобальный коэффициент средней степени соседних узлов, глобальный индекс дружбы и диаметр.

В качестве реальной сети рассматривается фрагмент сети Facebook с 10000 вершинами. Мы также строим граф на основе собранных данных из социальной сети Twitter. Результирующий граф содержит 2000000 вершин, у него считаются глобальные характеристики.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Выпускная квалификационная работа содержит введение, 6 глав, заключение, список использованных источников и 7 приложений.

В первой главе даются основные элементы теории вероятностей и графов. Концепции случайных экспериментов и событий, а также вероятности, исследуются в контексте стохастической устойчивости и равновероятности. Также дается подробное определение трех моделей случайных графов, а также локальных и глобальных характеристик, которые будут вычисляться для построенных графов и части социальной сети Facebook.

Случайный эксперимент — это модель ситуации, результаты которой непредсказуемы, хотя её параметры и условия остаются неизменными. Результаты этих экспериментов определяются случайными событиями, которые могут быть классифицированы как невозможные (никогда не происходят), достоверные (всегда происходят) или как-то между этими двумя категориями.

В контексте случайных событий важную роль играет понятие вероятности, числового значения, которое колеблется от 0 (для невозможных событий) до 1 (для достоверных событий). Особое внимание уделяется понятию равновероятных событий, которые все имеют одинаковую вероятность возникновения.

Вводится понятие классического вероятностного пространства как тройки элементов: пространство элементарных событий, множество всех возможных подмножеств этого пространства, и функция вероятности, определенная на этом множестве.

Также обсуждаются основные свойства вероятности, которые помогают анализировать совместные и взаимоисключающие события в рамках случайного эксперимента.

Определение схемы Бернулли обычно иллюстрируют на примере броска монеты и вероятности её падения на одну из двух сторон. Падение на решку обозначают как p и на орла как $q = 1 - p$, варианты приземления на ребро не учитывают. Проводится серия бросков монеты, где падение на решку отмечается как 1 и на орла как 0. N бросков монеты образуют схему испытаний Бернулли, и получаем случайную последовательность из нулей и единиц длиной N.

Вероятность того, что произойдет n конкретных событий, равна произведению их вероятностей $x_1, ..., x_n$. Так, для $\omega(x_1, ..., x_n)$ получаем $P(\omega) = p^{\sum_{i=1}^{n} x_i}\ q^{\sum_{i=1}^{n}(1-x_i)}$. Здесь можно видеть вероятностное пространство, элементами которого являются последовательности ω. Пространство элементарных событий состоит из возможных вариантов схемы Бернулли, и его мощность равна 2^n. Мощность множества событий равна 2^{2^n}, а вероятность равна сумме вероятностей элементарных событий. В схеме Бернулли вероятность наступления одного элементарного события равна $C_n^k p^k q^{n-k}$ и называется биномиальной по коэффициентам Ньютона.

Приводится описание схемы серий испытаний Бернулли, изображенной на рисунке. В ней рассматривают любую последовательность $n_1, n_2, ..., n_k$, где $n \in \mathbb{N}$. Для каждого i проводится ровно n_i испытаний Бернулли, и вероятность успеха в этом случае зависит от i и находится в пределах $[0, 1]$. Рассматриваемая последовательность натуральных чисел может быть бесконечной, но в контексте случайных графов чаще всего работают с конечными структурами, которые могут неограниченно расти.

Схема Бернулли аналогична серии бросков монеты, где шансы на решку и орла представлены p и $q = 1 - p$. Серия N бросков считается схемой испытаний Бернулли, создавая случайный набор нулей и единиц. Вероятность появления n конкретных событий равна их вероятностям, умноженным вместе. Пространство элементарных событий включает все возможные исходы схемы Бернулли.

Схема серий Бернулли рассматривает последовательность $n_1, n_2, ..., n_k$, где каждое число n_i отражает количество испытаний Бернулли, и вероятность успеха зависит от i.

Свойства пространства элементарных событий:

— Объединение любых двух событий является событием;

— Пересечение любых двух событий является событием;

— Дополнение любого события является событием;

— Весь набор результатов является событием;

— Условная вероятность исследует вероятность события A, учитывая, что событие B уже произошло.

Формула полной вероятности и формула Байеса позволяют обрабатывать условные вероятности, а события считаются взаимно независимыми, ес-

ли вероятность совместного события равна произведению их индивидуальных вероятностей.

Вероятностное пространство представляет собой модель возможных результатов эксперимента. Если пространство конечно, любая функция, переводящая его в набор вероятностей, является случайной величиной. Если нет, требуется, чтобы функция была измерима.

Функция распределения определяет вероятностное поведение данной величины. Случайные величины с такими функциями называются дискретными. Бесконечные пространства используют абсолютно непрерывные распределения, обеспечиваемые функцией-плотностью.

Случайные величины считаются независимыми, если это условие выполняется для всех пар событий.

Математическое ожидание — это средневзвешенное значение случайной величины, взвешенное по её вероятности. Главное свойство математического ожидания — линейность.

Дисперсия измеряет степень вариативности случайной величины, вычисляя среднее квадратичное отклонение от математического ожидания.

Случайный процесс — это функция времени, которая изменяется случайно и служит обобщением случайной величины. Результатом процесса является функция, а не число. Это может быть функция как одного, так и нескольких параметров. Результат каждого эксперимента представляет собой реализацию случайной функции.

Граф — это структура данных, которая состоит из множества вершин и ребер между ними. Графы бывают неориентированными, где ребра представляют собой неупорядоченные пары вершин, и ориентированными, где ребра — упорядоченные пары вершин.

Вершины называются смежными, если между ними есть ребро. Само ребро называется дугой, если оно представляет собой упорядоченную пару вершин. Если ребро соединяет вершину саму с собой, это называется петлей.

Простой граф — это граф без петель и мультиребер, то есть не содержит ребра, соединяющие одну и ту же пару вершин. Если такие ограничения не установлены, граф может быть мультиграфом или псевдографом.

Степенью вершины в графе называется количество ребер, которые исходят из этой вершины. Вершина может быть изолированной, то есть не иметь

связанных с ней ребер.

В некоторых случаях ребра графа могут быть взвешенными, то есть каждому ребру приписывается некоторое числовое значение или вес.

Вершины и ребра графа можно представить в виде матрицы смежности или списка смежности. В матрице смежности ребра между двумя вершинами обозначаются как единица, а отсутствие ребра — как ноль. Если граф разреженный, то есть имеет гораздо меньше ребер, чем вершин, то более эффективным способом представления может быть список смежности, который состоит из списков вершин, смежных с каждой вершиной.

Путь между вершинами a и b в графе определяется как последовательность вершин и ребер, начиная с a и заканчивая b. Длина пути равна количеству ребер в этой последовательности. Кратчайший путь — это путь с наименьшим количеством ребер между двумя вершинами. Путь считается простым, если все его вершины уникальны, за исключением возможно начальной и конечной.

Матрица кратчайших расстояний графа — это матрица, где каждый элемент $d[a][b]$ представляет собой кратчайшее расстояние между вершинами a и b.

Цикл — это путь, в котором начальная и конечная вершины одинаковы. Этот термин применяется только к неориентированным графам. Для ориентированных графов используется термин «контур». Простой цикл длиной 3 называется треугольником.

Полный граф — это граф, в котором каждая пара вершин соединена ребром. В отличие от этого, разреженный граф — это граф, в котором количество ребер примерно равно количеству вершин.

Граф считается связным, если существует путь между любой парой вершин. Если такого пути нет, граф считается несвязным. Максимальный подграф, все вершины которого связаны, называется компонентой связности.

Эксцентриситет вершины определяется как расстояние до самой отдаленной от нее вершины в графе. Радиус графа — это минимальный эксцентриситет среди всех вершин, а диаметр — максимальный. Вершина называется центральной, если ее эксцентриситет равен радиусу графа, и периферийной, если он равен диаметру.

Бесмасштабная сеть — это сеть, где степени вершин графа подчиняются

степенному закону.

Случайный граф — это термин, который охватывает множество графов, созданных на основе определенного вероятностного распределения или через случайный процесс. Теория случайных графов объединяет принципы теории вероятностей и теории графов, с основной целью исследования свойств и характеристик случайно сгенерированных графов.

Модель случайного графа представляет собой граф, который в каждый момент времени является случайным элементом из определенного набора графов, состоящих из определенного количества вершин. Детали модели определяются тем, как количество вершин зависит от времени, что входит в набор графов и распределением вероятностей этих графов.

Допустим, есть набор вершин для построения случайного графа. Рассматриваем все возможные рёбра, которые могут быть проведены между этими вершинами. Затем выбираем рёбра из этого набора на основе некоторого уровня вероятности. Если выбор ребра оказывается успешным, то оно добавляется в граф, в противном случае — нет.

Так строим случайный граф. Этот граф можно описать последовательностью чисел, где каждое число указывает на то, есть ли определённое ребро в графе или нет.

В альтернативном определении, известном как модель Эрдёша-Реньи, предполагается, что каждое ребро появляется в графе независимо от других, с определённой вероятностью. Когда исследуем свойства случайного графа, пытаемся определить набор графов, которые обладают этими свойствами, и изучаем вероятность этих свойств при увеличении числа вершин.

Исследователям в какой-то момент понадобилось создать модель графа всемирной паутины, то есть охарактеризовать особенности веб-графа, где вершинами служат различные ресурсы (вроде сайтов, статей или авторов). Дуга между двумя ресурсами проводится в том случае, когда один ресурс ссылается на другой. В таком графе можно иметь кратные дуги и даже петли, если ресурс ссылается на самого себя.

Учитывая эти особенности, применение модели Эрдёша-Реньи не подходит. Вместо этого, часто используются модели с приведенными ниже принципами, которые хорошо подходят для описания интернета, но могут быть использованы и для других типов сетей, например, социальных, транспорт-

ных или биологических.

При анализе веб-графа стоит учесть три основных аспекта: он обычно довольно разрежен, диаметр графа достаточно мал (от 5 до 7), а степень вершины зависит от суммы вероятностей.

Эти условия были описаны учеными Барабаши и Альбертом в их исследовании веба. Они особенно были заинтересованы в том, как появляются новые сайты в интернете. Когда появляется новый сайт, он обычно содержит ссылки на уже существующие сайты.

При этом вероятность наличия ссылки на старый сайт обычно тем выше, чем больше он популярен. Графы, в которых новые вершины добавляются с учетом этого принципа, обычно называют графами предпочтительного соединения.

У модели Барабаши-Альберта есть две проблемы. Первая — конечный граф зависит от начального графа. Если начальный граф был деревом, то даже с добавлением новых вершин он останется деревом. Если начальный граф был несвязным, то и конечный граф останется несвязным. Вторая проблема — случайный выбор вершин, который может привести к появлению треугольников в графе. Это стало поводом для создания следующей модели.

Модель Холма-Кима — это усовершенствованная версия другой модели, основанной на концепции предпочтительного присоединения. Эта модель была разработана Петтером Холмом и Бом Джун Кимом.

Они ввели идею локального коэффициента кластеризации для каждой вершины. В этом контексте «кластеризация» означает насколько плотно связаны вершины в графе.

Модель Холма-Кима представляет собой модификацию принципа предпочтительного присоединения и включает дополнительный шаг. Шаги алгоритма:

— Начальное состояние: граф начинается с определенным числом вершин, но без ребер;
— Рост: с течением времени в граф добавляется по одной вершине на каждом шаге. Каждая новая вершина привязана к некоторому числу ребер;
— Предпочтительное присоединение: каждое ребро новой вершины присоединяется к уже существующей вершине в графе.
— Вероятность присоединения зависит от степени, или числа ребер, уже

присоединенных к существующей вершине;

— Формирование триад: если на предыдущем шаге было добавлено ребро между двумя вершинами, тогда добавляется еще одно ребро, соединяющее новую вершину с случайным соседом уже связанной вершины.

Этот процесс создает структуру, которую называют безмасштабной сетью. Она имеет то же распределение степеней, что и другие подобные сети. Ученые обнаружили, что количество триад, или полностью связанных групп из трех вершин, увеличивается почти линейно с коэффициентом кластеризации.

Средний коэффициент кластеризации в сети это просто сумма локальных коэффициентов кластеризации всех вершин, деленная на общее количество вершин в графе.

Средняя степень соседних узлов вычисляется как сумма степеней всех смежных вершин, деленная на степень рассматриваемой вершины.

Глобальный коэффициент средней степени соседних узлов рассчитывается путем деления суммы средней степени соседних узлов всех вершин на общее количество вершин в графе.

Индекс дружбы это сумма степеней всех смежных вершин, деленная на квадрат степени рассматриваемой вершины.

Глобальный индекс дружбы определяется как сумма всех индексов дружбы вершин, деленная на общее число вершин в графе.

Во второй главе приводятся результаты вычислений глобальных характеристик графов, построенных на основе случайных моделей, а также демонстрируется динамика локальных характеристик на графиках. Также рассматриваются основные классы, реализующие программную логику для выполнения необходимых действий.

В рамках поставленной задачи предстоит научиться генерировать графы по моделям Эрдёша-Реньи, Барабаши-Альберта и Холма-Кима, а также вычислять их местные и глобальные особенности. Этот процесс осуществлялся на языке программирования Python 3 с помощью библиотек NetworkX для подготовки графа к визуализации и Matplolib для создания изображения.

Была реализована программная логика по формированию и основными действиями над графами. Закодирован алгоритм, который занимается подсчетом локальных и глобальных характеристик. Изображение полученного

графа демонстрируется на экране.

Также была рассмотрена сеть Facebook, извлеченная с использованием робота. Все данные были анонимизированы.

Для графа фрагмента Facebook были рассчитаны глобальные характеристики, по которым провели сравнение с графами на основе случайных моделей. Было заключено, что модель Холма-Кима оказалась наиболее близка к моделированию Facebook. Барабаши-Альберта показывает близкие значения средней степени вершин, в то время как модель Эрдёша-Реньи не сильно применима в области исследования графов социальных сетей.

В третьей главе даются основные сведения по платформе Twitter, которые будут полезны при разработке сборщика данных для этой социальной сети.

Twitter — одна из самых популярных социальных сетей, где пользователи обмениваются короткими сообщениями, называемыми твитами. Сеть была создана в 2006 году и стала неотъемлемой частью мировой коммуникации. Основной принцип работы Twitter — публикация коротких сообщений, которые могут содержать текст и мультимедийный контент. Пользователи могут подписываться на других пользователей, читать их твиты и взаимодействовать с ними. Twitter также используется для распространения новостей, мнений и событий.

Платформа предоставляет разработчикам доступ к данным через Twitter API, что позволяет создавать приложения для аналитики и мониторинга. Twitter имеет свои уникальные функции, такие как ретвиты, лайки и использование хештегов. Он доступен через веб-интерфейс и мобильные приложения для iOS и Android.

Twitter будет исследован и сравнен с графами, созданными на основе случайных моделей. Это позволит лучше понять, как формируются связи при появлении новых пользователей в системе, и появится возможность предсказывать дальнейший рост сети.

В четвертой главе рассказывается о создании системы, занимающейся сбором данных на основе ключевых слов в Twitter.

Для создания приложения, предназначенного для анализа данных в социальной сети Twitter по ключевым словам, был выбран язык программирования Python. Python является популярным инструментом для работы с

данными и обладает широкими возможностями для взаимодействия с API различных платформ.

Для работы с Twitter API в приложении была использована библиотека Tweepy, которая предоставляет удобный и интуитивно понятный интерфейс для работы с различными функциями API Twitter. С помощью Tweepy были реализованы функции для получения твитов по заданным ключевым словам, а также для сбора данных о пользователях и хештегах.

Полученные данные были анализированы с использованием различных библиотек Python, таких как Pandas, Matplotlib, Seaborn и другие. Эти библиотеки предоставляют мощные инструменты для обработки и визуализации данных, что позволило провести разнообразный анализ поведения пользователей, мнения общественности и других интересующих аспектов в Twitter.

Для создания пользовательского интерфейса приложения была использована библиотека Streamlit. Streamlit позволяет разработчикам создавать интерактивные веб-приложения на Python с минимальными усилиями. С его помощью был реализован удобный интерфейс, который позволяет пользователям вводить ключевые слова, выбирать период и настраивать параметры анализа данных.

Результаты анализа данных выводятся в виде диаграмм, графиков и других визуальных элементов, что облегчает восприятие и позволяет пользователям быстро и наглядно получить информацию о твитах, пользователях и хештегах, связанных с заданными ключевыми словами.

Для удобства разработки и управления зависимостями в проекте был использован менеджер пакетов pipenv. Он обеспечивает чистоту и стабильность проекта, а также позволяет легко установить и обновить необходимые пакеты и библиотеки.

Рассмотрен процесс настройки окружения на Windows 10 с использованием версии Python 3.10 и среды программирования PyCharm.

Далее подробно рассмотрен алгоритм действий при работе с программой. Данные выгружаются в популярном текстовом формате JSON. Можно регулировать, какие данные будут собраны, можно получить подписчиков автора сообщения, его подписки и много других параметров.

В пятой главе описан процесс разработки платформы, которая создает граф на основе собранных данных, процесс получения которых рассмотрен

в предыдущей главе. Граф визуализируется и представлен в такой форме, чтобы на нем можно было посчитать локальные и глобальные характеристики.

Было разработано приложение на языке программирования Python. Для взаимодействия с социальной сетью Twitter использовался программный интерфейс разработчика Twitter API v2.

В рамках разработки были использованы различные библиотеки для языка Python, устанавливаемые через менеджер пакетов pip. Среди основных таких библиотек можно отметить Tweepy, pandas, Json, random, docopt, pathlib и др. Для визуализации графа используются средства языка JavaScript.

Код также предусматривает обработку ошибок, связанных с достижением лимита на количество запросов в единицу времени и сбором всех доступных данных. На их основе строится граф, где вершинами выступают аккаунты пользователей, а ребро свидетельствует о наличии подписки между ними. Алгоритм осуществляет поиск, начиная с корневого аккаунта пользователя, и помечает все достижимые вершины одним цветом. Количество итераций алгоритма ограничено. По завершении поиска от одного популярного пользователя, запускается поиск от другого. Алгоритм был протестирован на графе с двумя миллионами вершин.

В шестой главе заново считаются локальные и глобальные характеристики для графов, построенных на основе случайных моделей. В этот раз подсчеты ведутся на графах с числом вершин до 2 млн. После этого глобальные характеристики полученных графов сравниваются с сетью Twitter, процесс формирования которой описан в прошлой главе.

Было определено, что модель Холма-Кима наиболее близка по глобальным характеристикам к рассмотренной части сети Twitter. Соответственно, именно ее рекомендуется применять при моделировании роста социальных сетей.

В приложениях представлены 7 листингов кода, содержащие реализацию алгоритмов случайных графов, логику по сбору данных Twitter, визуализацию графа и подсчет локальных и глобальных характеристик.

ЗАКЛЮЧЕНИЕ

В результате выполнения работы были достигнуты поставленные цели и выполнены основные задачи исследования.

В рамках работы был разработан программный код, способный генерировать графы на основе моделей случайных графов, включая модели Эрдёша-Реньи, Барабаши-Альберта и Холма-Кима. Был проведен анализ локальных и глобальных характеристик полученных сетей, что позволило лучше понять их структуру и свойства.

Также проведено сравнение глобальных характеристик фрагмента социальной сети Facebook с характеристиками графов, построенными с использованием упомянутых моделей. Было разработано приложение, занимающееся сбором данных в социальной сети Twitter, на основе которых был построен граф. Это позволило провести сравнение графов, созданных на основе случайных моделей, с фрагментом реальной сети.

По результатам исследования, модель Холма-Кима демонстрирует наиболее точное соответствие графам реальных сетей, поскольку ее характеристики более всего приближены к рассмотренным фрагментам реальных сетей. Эта модель может быть использована для прогнозирования развития социальных сетей и предсказания принципов, по которым они будут развиваться. Результаты эмпирического исследования, проведенного в рамках данной работы, подтверждают целесообразность применения модели Холма-Кима в аналитических исследованиях сетей и планировании инфраструктурных затрат для их поддержки.

Основные источники информации:

1 Боровков. А. А. Теория вероятностей / А. А. Боровков. — Москва: Издательство «Эдиториал УРСС», 1999.

2 Гнеденко. Б. В. Курс теории вероятностей / Б. В. Гнеденко. — Москва: Издательство «Эдиториал УРСС», 2005.

3 Райгородский. А. М. Модели случайных графов / А. М. Райгородский. — Москва: Издательство МЦНМО, 2011.

4 Райгородский. А. М. Модели интернета / А. М. Райгородский. — Долгопрудный: Издательский дом «Интеллект», 2013 — 64 с.

5 Erdos, P. On random graphs / P. Erdos, A. Renyi // Publicationes Mathematicae Debrecen. — 1959. — Vol. 6. — Pp. 290-297.

6 Barabasi, A.-L. Emergence of scaling in random networks / A.-L. Barabasi, R. Albert // Science (New York, N.Y.). — 1999. — Vol. 286, no. 5439. — Pp. 509-512.

7 Empirical validation of the buckley-osthus model for the web host graph: Degree and edge distributions / M. Zhukovskiy, D. Vinogradov, Y. Pritykin, L. Ostroumova, E. Grechnikov, G. Gusev, P. Serdyukov, A. M. Raigorodskii // CoRR. — 2012.

8 Шапошников, К. С. Генерация сложных сетевых структур на основе оптимизированной модели с предпочтительным присоединением / К. С. Шапошников, И. Д. Сагаева, С. П. Сидоров // Информационные технологии и математическое моделирование (ИТММ-2019): Материалы XVIII Международной конференции имени А.Ф. Терпугова. — 2019. — Vol. 2. — Pp. 75-79.

9 Twitter API Documentation | Docs | Twitter Developer Platform [Электронный ресурс]. — URL: https://developer.twitter.com/en/docs/twitter-api (Дата обращения 16.05.2023). Загл. с экр. Яз. англ.

10 The Python Standard Library — Python 3.10.10 documentation [Электронный ресурс]. — URL: https://developer.twitter.com/en/docs/twitter-api (Дата обращения 16.05.2023). Загл. с экр. Яз. англ.