МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**
Кафедра математической кибернетики и компьютерных наук

АНАЛИЗ ХАРАКТЕРИСТИК МОДЕЛЕЙ СОЦИАЛЬНЫХ СЕТЕЙ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента 2 курса 273 группы
направления 02.04.03 — Математическое обеспечение и администрирование
информационных систем
факультета КНиИТ
Есина Ильи Дмитриевича

Научный руководитель
зав. каф., к. ф.-м. н., доцент _____ С. В. Миронов

Заведующий кафедрой
к. ф.-м. н., доцент _____ С. В. Миронов

Саратов 2025

ВВЕДЕНИЕ

Актуальность темы. Изучение социальных, биологических процессов, зачастую представленных в виде системы из набора узлов, требует наличие моделей, позволяющих описывать сложные, естественные процессы. Однако модели не всегда могут гарантировать точность полученных результатов.

Цель магистерской работы — проведение эмпирических исследования существующих моделей социальных графов и сравнение полученных результатов с данными, полученными в реальных социальных сетях.

Поставленная цель определила **следующие задачи:**

- Изучить работы на похожие темы и сделать выводы.
- Изучить актуальные модели построения безмасштабных сетей.
- Изучить методы и инструменты для анализа социальных сетей.
- Найти данные для исследования.
- Изучить возможности сбора данных, подготовить пространство для сбора данных. Осуществить сбор данных.
- Выбрать ключевые метрики.
- Предложить собственную модель построения безмасштабных сетей.
- Проанализировать актуальные методы кластеризации, которые могут быть использованы при работе с графами.
- Сравнить результаты, полученные в ходе анализа моделей с результатами, полученными в реальных сетях.

Практическая значимость магистерской работы Практическая значимость исследования заключается в изучении и оптимизации структуры безмасштабных сетей, что позволяет повысить эффективность распространения информации и устойчивость к целевым атакам. Предложенная модель может быть применена для улучшения рекомендательных систем и минимизации каскадных отказов в социальных платформах.

Структура и объём работы. Магистерская работа состоит из введения, 3 разделов, заключения, списка использованных источников и 2 приложений. Общий объем работы — 69 страниц, из них 63 страницы основное содержание, включая 26 рисунков и 3 таблицы, список использованных источников информации — 27 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Постановка задачи, анализ существующих решений» посвящен:

1. Обзору наиболее популярных существующих исследований и их ответов на различные вопросы.
2. Выбору метрик для исследования.
3. Анализу методов машинного обучения.
4. Изучению моделей растущих сетей.
5. Изучению и подготовке необходимого инструментария.

Перед проведением исследования был проведен анализ существующих работ. Так например, в одной из работ анализировались метрики социальных сетей, такие как центральность, ассортативность, плотность и коэффициент кластеризации. Были рассмотрены характеристики сетей, разделенных на кластеры. В качестве программного инструмента использовался Python с библиотекой NetworkX. В заключении автор пришел к выводу, что социальные сети имеют ограниченный набор данных о социуме, не могут предсказывать его поведение и делать предпочтения на основе этой информации.

Следующая работа была посвящена анализу кластеризации библиометрических сетей и спектральному анализу сети цитирования научных журналов. В исследовании было проведено сравнение алгоритмов WTR и LEV, где WTR показал лучшие результаты в кластеризации библиометрической сети. Это подтверждает важность спектральных методов для эффективной кластеризации научных сетей.

В работе были изучены классические модели построения безмасштабных сетей.

— Модель Барабаши—Альберт является одной из первых моделей генерации случайных, безмасштабных сетей с использованием принципа предпочтительного соединения.

— В основе триадного замыкания лежит свойство 3 вершин A, B, C, которое заключатся в том, что если существуют связи $A-B$ и $B-C$, то существует тенденция к формированию новой связи $B-C$.

Принцип предпочтительного соединения — это ключевая идея, лежащая в основе модели Барабаши—Альберт и триадного замыкания для построения случайных сетей. Этот принцип объясняет, что вероятность присоединения кон-

кретной вершины реб- ром к новой вершине пропорциональна степени данной вершины.

При изучении существующих моделей, было выявлено упрощенное отображение действительности. Модель Барабаши—Альберт и модель триадного замыкания при добавлении новых вершин демонстрируют выраженную центростремительную динамику. Это отображается на группе вершин, которые были добавлены в граф на начальных итерациях алгоритма. Данная группа вершин де- монстрирует завышенные коэффициенты центральности и кластеризации, что подтверждается практическими наблюдениями.

Была предложена собственная модель построения безмасштабных сетей основанная на принципе из A кластеров.

В работе рассматриваются сети, изменяющиеся в дискретном времени $t = 0, 1, 2....$

Состояние растущей сети в момент времени t будем описывать графом $G(t) = (V(t), E(t))$.

1. s — изначальное количество кластеров размера m_1.
2. p — вероятность появления нового кластера, где $p \in [0, 1]$.
3. m — количество рёбер, которые нужно добавить вершине u_t в момент времени $t, \quad m < m_1$.
4. m_1 — количество вершин, добавляемых в момент создания кластера.
5. σ — среднеквадратическое отклонение для значения m_1.
6. p_2 — вероятность появления ребра к соседнему кластеру, где $p_2 \in [0, 1]$.
7. T — количество итераций алгоритма.

Данная модель будет строить граф с разрастающимся количеством центров вершин с большой плотностью.

В момент времени t_0 необходимо создать s групп вершин, каждая из которых содержит m_1 вершин. Пусть $A = \{a_1, a_2, ...a_n\}$ множество всех кластеров. При этом $a_i = \{u_1, u_2, \dots u_k\}$ - множество, которое содержит вершины принадлежащие данному кластеру, тогда $|A|$ - количество этих кластеров, $|a_i|$ - количество вершин в данном множестве.

В каждый последующий момент времени t:

1. С вероятностью p_1 добавляем новый кластер a_t из
   $n \sim \mathcal{N}(m_1, \sigma^2), \quad m_1 \in \mathbb{N},\ m_1 > 0$ вершин. Данный кластер образует полный граф.

2. С вероятностью 1 - p добавляем вершину u_t, при этом:
   - До момента добавления рёбер происходит выбор принадлежности вершины к a_i кластеру по методу предпочтительного соединения.
   - Первое ребро от вершины u_t добавляется к одной из вершин $u_j \in a_i$ по методу предпочтительного соединения.
   - Каждое из последующих $m - 1$ ребер добавляется к вершине u_t по следующим правилам:
     - С вероятностью $1 - p_2$ добавляем ребро между вершиной u_t и вершиной $u_i \in a_i$.
     - С вероятностью p_2 добавляем ребро между вершиной u_t и вершиной $u_i \notin a_i$.

Для изучения структуры графа были выбраны следующие метрики:

- Метрики, связанные со структурой графа. Данные метрики описывают основные характеристики графа как целой сущности. К ним можно отнести:
  1. Количество вершин.
  2. Количество рёбер.
  3. Плотность графа.
  4. Диаметр.
  5. Радиус.
- Метрики центральности. Данные метрики показывают, насколько централизована, важна каждая вершина в графе. К ним можно отнести:
  1. Степень вершины.
  2. Средняя степень.
  3. Центральность по степени.
  4. Центральность по кратчайшему пути.
  5. Центральность по промежуточным путям.
  6. Эгоцентрическая центральность.
- Метрики связности. Данные метрики показывают, насколько связан граф. К таким можно отнести:
  1. Коэффициент связности.
  2. Компоненты связности.
  3. Коэффициент кластеризации.
- Метрики путей. Данные метрики фокусируются на анализе расстояний между вершинами графа.

1. Средняя длина кратчайшего пути.
2. Кратчайшие пути.

Анализируя инструменты было решено использовать язык Python и библиотеку NetworkX. NetworkX —– это универсальная библиотека для создания, манипуляции и изучения структуры, динамики и функций сложных сетей. Она написана на языке Python и предоставляет широкий набор инструментов для анализа графов.

Помимо исследования метрик графа необходимо было рассмотреть и методы машинного обучения которые позволяют решать большое количество задач. При анализе графовых структур большой пласт задач решается с помощью методов машинного обучения. Все задачи, решаемые с помощью данных методов, можно поделить на 3 основных категории:

— Классификация. В данный список попадают задачи, в которых необходимо предсказывать значение для каждой вершины по отдельности. Это может быть полезно, например, для предсказания роли узлов в социальных сетях (например, идентификация пользователей с высоким социальным статусом). Помимо предсказания меток для вершин графа, можно предсказывать вероятности для ребер графа. Например, в социальных сетях можно предсказать, кто может стать другом пользователя. Для решения такого рода задач могут подойти, как различного рода нейронные сети, так и методы логистической регрессии.
— Кластеризация графа. В данный список попадают задачи на выявление группы узлов, которые более плотно связаны между собой, как на основе сторонних признаков (например, возраст, пол), так и на основе данных отдельно взятой вершины (например, степень вершины, коэффициент кластеризации, индекс дружбы). Для изучения кластеризации графа следует обратить внимание на алгоритмы кластеризации, такие как: DBSCAN, HDBSCAN, KMeans, Spectral Clustering и другие. Данные алгоритмы могут помочь, как в решении задач кластеризации, так и в обнаружении аномалий.
— Предсказание временных изменений в графах. Задачи предсказания помогают ответить на большой список вопросов. Один из таких вопросов формулируется как оценка вероятности появления новых вершин вершин и ребер в графе. По мимо этого можно строить предположения о том, как

будут меняться уже существующие связи в графе. Для этого можно использовать рекуррентные нейронные сети или методы на основе эволюции графов.

Второй раздел «Построение модели» посвящен реализации программных элементов для изучения полученных графов.

Для удобной работы с графами, их обработкой, а так же последующей визуализацией необходимы современные инструменты. Язык Python был выбран ввиду большого списка инструментов, библиотек для работы с большими объемами данных. Использовались следующие инструменты:

1. NetworkX — библиотека, предназначенная для создания, манипулирования и анализа графов. Данная библиотека предоставляет удобный интерфейс для работы с графами, реализует множество алгоритмов для их анализа. Имеет встроенную поддержку генерации графов по модели Барабаши—Альберт.
2. Matplotlib — библиотека для визуализации данных. Модуль pyplot предоставляет интерфейс для создания графиков и диаграмм. Это основополагающий инструмент для визуализации анализа.
3. Pandas — библиотека для обработки и анализа данных. Представляет собой удобный инструментарий для работы с большими массивами данных, их анализом и группировкой. Отлично совмещается с NetworkX и Matplotlib.
4. Seaborn — библиотека для визуализации данных, построенная на базе Matplotlib. Будем использовать для построения некоторых видов графиков, которые отсутствуют в библиотеке Matplotlib.
5. Community предоставляет функции для обнаружения сообществ в графах, используя метод Алгоритм Лувена (Louvain Algorithm), который направлен на поиск наиболее плотных кластеров в графах.
6. Multiprocessing позволяет создавать многозадачные (параллельные) программы в Python, используя несколько процессов, что позволяет эффективно использовать многопроцессорные системы. Данный модуль очень помогает при работе с большими объемами данных.

Изначально большая часть системы строилась в IDE PyCharm. Через некоторое время было принято решение перейти на Jupyter Notebook. У данного решения есть один фактор, который позволил сократить время на выполнение

определенных участков кода. Jupyter позволяет запускать ячейки с кодом поочередно, что делает процесс разработки и исследования более гибким и удобным. Можно выполнять код по частям, проверять результаты, изменять их и сразу видеть изменения. Это очень удобно в случае, когда процесс создания графа является долгим, а метод по поиску метрики более быстрым. Не имеет смысла создавать один и тот же граф несколько раз, затрачивая большой объем времени. Для перехода на Jupyter Notebook был поднят Jupyter Notebook server. Для реализации алгоритмов по выявлению сообществ и последующего анализа необходимо реализовать систему, которая позволит реализовать алгоритмы кластеризации, а так же визуализации полученных результатов. Так как заранее не предоставлялась возможность оценить, какой из описанных ранее алгоритмов кластеризации графов будет оптимальным, необходимо было подготовить реализацию всех подходящих алгоритмов и исследовать полученные результаты.

Для проведения анализа социальных графов был подготовлен датасет с использованием VK API, который позволил собрать данные о взаимодействиях между пользователями социальной сети ВКонтакте. С помощью данного API были извлечены данные о друзьях пользователей, их взаимосвязях, а также дополнительная информация о группах, интересах и активностях в сети. Эти данные стали основой для построения графов, которые использовались для изучения различных метрик, таких как центральность, коэффициент кластеризации и другие параметры, характерные для социальных сетей. Полученный датасет был использован для анализа реальных социальных сетей и дальнейшего сравнения с теоретическими моделями, что позволило получить более точное представление о структуре и поведении реальных социальных графов.

Третий раздел «Анализ результатов» посвящен изучению полученных результатов исследования. В первой части исследования рассматривалась модель Барабаши—Альберт. Были построены графы с различным количеством вершин и параметром m. В ходе анализа выделялись ключевые характеристики графов, такие как распределение коэффициентов кластеризации, степени вершин, а также центральности. Для графа с 1000 вершинами и m=10 было выявлено, что большинство вершин имели коэффициенты кластеризации, близкие к нулю, что свидетельствовало о разреженной структуре. Однако несколько вершин имели более высокие коэффициенты, что указывало на наличие небольших

сообществ. Распределение степеней показало наличие вершин с высокими степенями и множества вершин с низкими степенями. Центральности, включая степень, посредничество и собственные векторы, продемонстрировали схожие результаты, где несколько ключевых вершин играли важную роль в сети.

Когда параметры модели изменялись, наблюдались изменения в распределении степеней: в новом графе распределение стало более сглаженным, с меньшим количеством выбросов. Эти результаты подтвердили зависимость характеристик графов от гиперпараметров модели.

Затем анализировались графы, построенные по модели Эрдёша—Реньи, с различными значениями параметра p, который отвечал за вероятность появления ребра между вершинами. Для различных значений p были построены графы с 5000 вершинами, и результаты показали, что при изменении p наблюдаются изменения в коэффициентах кластеризации, однако метрики центральности оставались достаточно стабильными. На графиках были замечены высокие степени кластеризации, и распределение степеней вершин показало, что большинство вершин имеют низкие степени, а несколько вершин — высокие. В целом, эти графы отличались от графов модели Барабаши—Альберт в плане структуры и распределений.

Для анализа моделей триадного замыкания было создано множество графов с различными значениями гиперпараметров, таких как p и m. Результаты показали, что коэффициенты кластеризации были сосредоточены в диапазоне 0.2–0.5, что указывало на наличие кластеров. Распределение степеней было также степенным, что соответствовало характеристикам модели Барабаши—Альберт. Для графов с большим количеством вершин было показано, что степень кластеризации и распределение степеней изменялись в зависимости от параметров модели.

В дополнение к математическим моделям были рассмотрены графы, полученные из реальных социальных сетей, таких как Google+, AnyBeat и Advogato. Результаты показали, что реальные графы имели схожие характеристики с графами, построенными по модели Барабаши—Альберт. Однако по параметрам, таким как ассортативность, наблюдались различия. Это указывало на то, что реальные данные имеют свои особенности, отличные от теоретических моделей.

В заключение была предпринята попытка кластеризации графов с использованием алгоритмов DBSCAN и спектральной кластеризации. Алгоритм

DBSCAN не дал удовлетворительных результатов, так как всегда образовывался единственный кластер, включающий все вершины. Однако спектральная кластеризация продемонстрировала более интересные результаты, особенно для графов, построенных по модели Барабаши—Альберт, где кластеры имели явные отличия по меткам.

ЗАКЛЮЧЕНИЕ

В данной работе были анализированы и изучены различные социальные сети, определены целевые метрики для снятия показателей при исследовании социальны сетей. Рассмотрено большое количество работ по данному вопросу, изучены и проанализированы работы на схожие тематики. Достигнуто понимание в работе модели Барабаши—Альберт для построения социальных графов а так же получен доступ и приобретен опыт работы с API VK.

В ходе исследования графов построенных по модели Барабаши—Альберт, триадного замыкания, Эрдёша—Реньи, а так же по графам, которые были получены на основе реальных социальных сетей различного рода можно сделать вывод о том, что в большинстве случаев, на изучаемых метриках, которые описывались в ходе выполнения работы, графы имеют близкие по значению показатели. Исходя из этого можно предположить, что графы, построенные по модели Барабаши—Альберт, триадного замыкания, Эрдёша—Реньи, подходят для изучения социальных графов, в случаях если получение данных из реальных социальных сетей не предоставляется возможным.

В ходе изучения кластеризации графов построенных с помощью алгоритмов генерации случайных безмасштабных сетей можно сделать вывод о том, данные графы плохо поддаются кластеризации в не зависимости от выбранных параметров модели и алгоритмов кластеризации.

В ходе работы, предложена собственная модель построения социальных графов, основанная на методе динамического добавления кластеров. Данная модель имеет существенные отличия и предлагает альтернативный вариант построения графов, имитирующих социальные процессы. Главным отличием данной модели является децентрализация групп вершин с высокой степенью по всей поверхности графа, что позволяет проводить кластеризацию аналогично тому, как бы проводилась кластеризация на данных полученных из реальных источников.

Основные результаты исследования были опубликованы в статье: «Модель формирования растущей сети с предпочтительным присоединением к q кластерам» в научном журнале «Студенческий вестник».