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 ТЕЛЕФОНИИ УНИВЕРСИТЕТА

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 451 группы
направления 09.03.04 — Программная инженерия
факультета КНиИТ
Сарафанникова Даниила Александровича

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

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

Саратов 2026

В последние десятилетия анализ сложных сетей стал одним из важных направлений исследований. Социальные сети, транспортные системы, биологические системы, интернет-инфраструктура могут быть представлены в виде графов, свойства которых позволяют выявлять закономерности структурной организации и функционирования подобных систем.

Особый интерес представляют телефонные сети, поскольку они отражают коммуникации между людьми и обладают как структурными, так и динамическими характеристиками, образуя масштабную сеть, развивающуюся нетривиальным образом во времени. Для проведения анализа конкретной сети и прогнозирования ее поведения в будущем, а также для изучения применимости различных теоретических моделей сложных сетей необходимо иметь датасеты, собранные на основе настоящих звонков. Одной из проблем является ограниченность множества доступных в открытом доступе датасетов. Ввиду требований конфиденциальности и особенностей сбора данных подготовка такого датасета — достаточно важная задача.

В работе рассматривается подготовка и анализ датасета телефонных звонков абонентов СГУ. На основе собранных данных строится граф взаимодействий и исследуются его свойства. Полученные результаты могут быть использованы как для дальнейшего изучения телефонии СГУ, так и для сопоставления с теоретическими моделями сложных сетей.

Цель работы: подготовка и анализ датасета телефонии университета на основе собранных данных.

Для достижения данной цели поставлены следующие задачи:

- изучить теорию о сложных сетях и популярных моделях;
- изучить существующие датасеты;
- очистить и унифицировать собранные данные;
- провести анализ структурных и динамических характеристик графа взаимодействий внутри университета;
- подготовить датасет к публикации.

Актуальность данной работы обусловлена несколькими факторами:

- дефицит реальных открытых телекоммуникационных данных;
- потребность в наличии датасетов для проверки теоретических моделей;
- практическая ценность анализа для изучения поведения сети телефонии СГУ и прогнозирования;

— использование современных методов анализа графов.

Таким образом, работа является актуальной не только в рамках конкретной телефонной сети, но и в рамках изучения сложных сетей в целом.

Бакалаврская работа состоит из введения, 3 разделов, заключения, списка использованных источников и 4 приложений. Общий объем работы — 58 страниц, из них 44 страницы — основное содержание, включая 20 рисунков и 1 таблицу. Список использованных источников информации состоит из 28 наименований.

В первом разделе «Телефония как сложная сеть» рассмотрены роль датасетов звонков в изучении свойств сложных сетей, существующие проприетарные и публичные датасеты с приведением результатов, полученных на их основе. Детальные записи о звонках (CDR) — один из наиболее ценных источников эмпирических данных: они позволяют изучать реальную структуру некоторого общества, взаимодействия в нем, процессы передачи информации, поведение людей в масштабе организаций, сообществ, городов и даже стран. Иначе говоря, речь идет и о структурных характеристиках, и о динамических. Изучение сетей телефонии позволяет решать ряд задач: от классических (распределение степеней, коэффициенты кластеризации, выделение сообществ) до прикладных (предсказание оттока клиентов и дефолтов, выявление точек притяжения и перегруженных участков сетей передачи информации). Ценность датасетов определяется размером, большой выборкой абонентов, несколькими источниками (не только телефонные звонки, но и SMS и интернет-сессии), дополнительной информацией об абонентах (местоположение, возраст и проч.).

Среди проприетарных датасетов рассмотрены 4 штуки. Полученные сети отличаются большими размерами (порядка 1–10 миллионов связей). Основные результаты включают в себя неприменимость степенного закона распределения степеней (вместо него может быть степенной закон с экспоненциальной отсечкой), подтверждение гипотезы слабых связей Грановеттера, наблюдение эффекта малого мира, низкие коэффициенты кластеризации, применимость степенного с экспоненциальной отсечкой закона для распределения длительностей звонков. Из нестандартных методов изучения выделяются: построение статистически валидированных сетей для очистки от обзвонов и горячих линий, применение характеристик графа для обучения ML-моделей, эпидемологическое моделирование, выделение городских точек притяжения специальными

алгоритмами.

Среди публичных датасетов аналогично рассмотрено 4 штуки. Датасет Reality Mining отличается малыми размерами (100 абонентов), слабой обобщаемостью результатов (абоненты — студенты и сотрудники лаборатории), несколькими источниками данных (звонки, смс, физическая близость). D4D отличается большими размерами (5 миллионов звонков) и информацией о местоположении абонентов, но данные представлены в 4 форматах, ни один из которых не позволяет построить полноценную сеть. Датасет, собранный в Милане и Торонто представляет из себя большой объем информации, привязанный к пространственной сетке и агрегированный по сетке времени с 10-минутным интервалом. Датасет Nodobo является развитием идеи Reality Mining с увеличенным размером выборки, но все еще отличается слабой обобщаемостью.

В заключении данного раздела сделан вывод о наличии ограничений и компромиссов в открытых датасетах. В свою очередь проприетарные датасеты отличаются большей выборкой и разнообразием данных, что позволяет получать значимые научные результаты.

Во втором разделе «Модели формирования сложных сетей» рассмотрены модели Эрдеша-Реньи (пуассоновское распределение степеней, наличие порога связности), Барабаши-Альберт (механизм предпочтительного присоединения, степенной закон с $\gamma = 3$), Холма-Кима (аналогично Барабаши-Альберт, но за счет триадного замыкания повышается кластеризация) и Stochastic Block Model (выделение сообществ, современные модификации). Данный раздел приведен для возможности сравнения полученных далее результатов с популярными теоретическими моделями, а также для изучения структуры сети.

Третий раздел «Подготовка и анализ датасета» включает в себя описание проведенной работы над базой звонков и абонентов СГУ. Он состоит из 8 подразделов.

Подраздел «Используемые технологии» описывает выбор языка программирования и библиотек с обоснованием. Для проведения анализа выбран Python 3.12. Среди используемых библиотек: Pandas (работа с датасетом, проведение предобработки), NetworkX (построение графов и изучение их свойств), powerlaw (проверка применимости законов распределения), graph-tool (применение Stochastic Block Model), BeautifulSoup (сбор дополнительной информации с сайта СГУ). Исследования велись в Jupyter Notebook для удобства структу-

рированного представления кода, текстового сопровождения и визуализаций в одном месте.

Подраздел «Очистка и подготовка данных» описывает методику обработки абонентской базы и базы звонков (унификация номеров, удаление некоторых звонков, переход от внешних номеров к внутренним). В итоге из 534323 звонков были оставлены 302201. База абонентов состоит из 467 элементов. В рамках работы проведен анализ для звонков между абонентами СГУ (24114 звонков при удалении 866 звонков без времени начала).

В подразделе «Исследование свойств графа» имеются 4 пункта. В первом рассмотрено построение графов: направленного мультиграфа звонков $\overrightarrow{G}_m$ (208 вершин, 24114 ребер), взвешенный орграф связей $\overrightarrow{G}_w$ (вес ребра — количество звонков между парой абонентов, 208 вершин, 2534 ребер, плотность 0,06), взвешенный ненаправленный граф связей G_{w} (208 вершин, 1991 ребро, плотность 0,09). Графы $\overrightarrow{G}_m$ и $\overrightarrow{G}_w$ являются слабо связанными, но не сильно связанными: 32 компоненты сильной связности, из них 31 компонента состоит только из 1 абонента (абоненты, которые только звонили или которым только звонили).

Во втором пункте изучается распределение степеней: например, в $\overrightarrow{G}_m$ средняя степень входа составляет 115,93, среднее квадратичное отклонение степени входа 249,62, медиана 30, максимальная степень 2695. Средняя степень выхода аналогично 115,93, СКО степени выхода 202,26, медиана 30, максимальная степень 1313. Для распределения степеней вершин в мультиграфе были опробованы:

— степенной закон $d(x) \sim x^{-\alpha}$;
— логнормальный закон $d(x) \sim x^{-1}e^{-(\log(x)-\mu)^2/2\sigma^2}$;
— степенной с экспоненциальной отсечкой $d(x) \sim x^{-\alpha}e^{-\lambda x}$;
— растянутый экспоненциальный $d(x) \sim x^{\beta-1}e^{-(\lambda x)^\beta}$.

Визуализация представлена на PDF ($P(X = x)$) и на CCDF (complementary cumulative distribution function, $P(X \geq x)$) в лог-лог масштабе. Использование CCDF в подобных задачах обусловлено тем, что часто в хвостах распределения степеней наблюдается шумная область, за счет интегральной природы CCDF этот шум сглаживается. Логарифмический масштаб удобен тем, что степенной закон в нем отображается в виде прямой. Для удобства была реализована функция `analyze_laws`, которая позволяет сравнить распределения между собой и отобразить их на графике. Также было проведено сравнение применимости

законов на основе логарифмического отношения правдоподобий и p-value.

В случае с $\overrightarrow{G}_m$ данные наилучшим образом описываются степенным законом с экспоненциальной отсечкой, $\alpha = 0.72, \lambda = 0.0014$. В случае сравнения применимости степенного закона и логнормального логарифмическое отношение правдоподобий R составляет -80.95, притом значимость (p-value) этого отношения порядка 10^{-13}, что однозначно свидетельствует о неприменимости степенного закона здесь.

В случае с $\overrightarrow{G}_w$ данные описываются как степенным законом с отсечкой ($\alpha = 0.28, \lambda = 0.0313$), так и растянутым экспоненциальным законом ($\lambda = 0.0462, \beta = 0.85$). Сравнение степенного закона с логнормальным аналогично свидетельствует о неприменимости степенного закона ($R = -98.08, p = 1.98 \cdot 10^{-15}$). Выбор между степенным законом с отсечкой и растянутым экспоненциальным неоднозначен ($R = 0.72, p = 0.23$) ввиду большого значения p-value.

Данный результат, а именно плохая подгонка данных степенным законом, который теоретически получен для моделей Барабаши-Альберт и Холма-Кима, может быть связан с тем, что механизм предпочтительного присоединения не может быть применен без модификаций в реальности, так как телефонные абоненты не могут совершать слишком много звонков. Другое возможное объяснение — недостаточное количество данных в хвосте, возможно, при большей выборке данных распределение степеней могло бы быть лучше описано степенным законом. Однако стоит учитывать, что даже для датасетов больших размеров подгонка распределения степеней лучше всего осуществлялась не степенным законом.

В третьем пункте рассмотрены метрики центральности вершин: по степени, по собственному вектору, по близости, по посредничеству, по потоку посредничества и потоку близости, communicability betweenness centrality, PageRank, HITS. После вычисления каждой метрики для каждой вершины, вершины могут быть пронумерованы в отсортированном порядке согласно их центральности по данной метрике, то есть каждая вершина получает некоторый ранк в данной метрике. В результате усреднения ранков по всем метрикам вершины были отсортированы. Наиболее центральные вершины соответствуют техническим подразделениям университета, а также деканатам факультета.

В заключительном пункте подраздела рассмотрены коэффициенты кла-

стеризации, введенные для описания склонности вершин к группировке. Существуют локальный коэффициент кластеризации, определенный для каждой вершины, и глобальный (транзитивность). Оба коэффициента основаны на понятии треугольников в графе. Под треугольником понимают клику размером 3. Определение локального коэффициента кластеризации может быть обобщено для взвешенных графов и ориентированных. Также для описания общего уровня кластеризации в сети предлагается считать средний коэффициент кластеризации. В простейшем случае, граф G_{w} без учета весов, средний локальный коэффициент кластеризации составляет 0.42, транзитивность 0.31. В случае учета направлений, но без учета весов (орграф связей $\overrightarrow{G}_w$) средний коэффициент кластеризации составляет 0.32. В случае учета и направлений, и весов (кол-во звонков), то есть при рассмотрении графа $\overrightarrow{G}_w$, средний коэффициент кластеризации составляет $0.23 \cdot 10^{-2}$.

В следующем подразделе «Статистика звонков» рассмотрено распределение длительности звонков с представлением гистограммы в логарифмическом масштабе. Как и с анализом распределения степеней вершин, рассмотрен ряд законов. Оказалось, что данные могут быть хорошо описаны логнормальным законом с параметрами $\mu = 3.58, \sigma = 1.14$. Обычно логнормальный закон возникает при мультипликативной природе случайной величины, то есть длительность разговора может быть описана комбинацией множества факторов, притом факторы взаимодействуют путем перемножения. Хвост у логнормального распределения менее «тяжелый», чем у степенного распределения, что в данном случае соотносится с тем фактом, что экстремально длинные звонки крайне маловероятны.

В рамках этого же подраздела рассмотрено распределение длительностей интервалов между звонками. Рассмотрим пару абонентов А и Б. Пусть А позвонил Б. Зададимся вопросами: через сколько А снова позвонит Б, через сколько А позвонит Б или Б позвонит А, через сколько Б позвонит А. Для всех вопросов рассмотрим распределение интервалов в логарифмическом масштабе и в линейном масштабе для первых 100 часов. Во всех трех случаях замечена затухающая суточная сезонность, заметная на графиках в линейном масштабе.

В следующем подразделе «Темпоральные графы» изучена динамика графов с использованием двух методов. Первый метод — метод скользящего окна. Будем рассматривать подграфы звонков, совершенных в период из некоторого

количества последовательных дней. Окна из дней будем сдвигать на некоторый шаг. Движение притом возможно как по календарю из всех дней, так и по календарю из только рабочих дней (без учета выходных дней и официальных праздников согласно производственному календарю РФ пятидневной рабочей недели). Для каждого графа возможно построить CCDF ($(P(X \geq x)$) распределения степеней, далее эти CCDF можно сравнить по некоторой метрике (L2 или Kolmogorov-Smirnov). При построении и сравнении CCDF стоит учитывать, что функция ступенчатая. Если в некоторой точке «нет» значения (например, есть вершины степени 4 и вершины степени 6, но нет вершин степени 5), то из определения CCDF следует, что в точке 5 берется значение из точки 6. Сравнение CCDF возможно как относительно усредненной функции, так и относительно функции из первого окна. Окна с малым количеством звонков (меньше 100) отбрасывались. Для удобного анализа графиков для разных параметров была реализована функция `slide_window`. Были замечены структурные изменения в летний период, они могут быть связаны не с работой приемных комиссий, а скорее с уменьшением взаимодействия подразделений в период отпусков (т.к. здесь рассматривается граф звонков внутри СГУ, а активность приемной комиссии направлена вне СГУ). Изменения, начало которых наблюдается в середине 2025 года, могут быть связаны с блокировкой звонков в популярных мессенджерах.

Второй метод — метод с постоянным количеством ребер. Будем итерироваться по отсортированном по времени списку звонков и поддерживать мультиграф звонков с постоянным количеством звонков, удаляя старые звонки и добавляя новые. Для каждого момента времени можем определить количество уникальных абонентов, задействованных в текущем графе. Полученные результаты в целом согласуются с динамикой, показанной в методе скользящего окна. Аналогично можно поступить с неориентированным графом связей G_{w}: поддерживаем некоторое количество актуальных связей (обновление актуальности связи происходит при звонке в данной паре абонентов).

Следующий подраздел — «Стохастическая блочная модель». Для изучения применимости Stochastic Block Model (SBM) к данной сети будем использовать функционал библиотеки `graph-tool`. Прежде всего необходимо преобразовать мультиграф звонков $\overrightarrow{G}_m$, хранимого в формате библиотеки `NetworkX`. В `graph-tool` отсутствует поддержка мультиграфов, поэтому мы будем строить ориентированный взвешенный граф, агрегируя ребра по парам вершин. В

качестве веса мы будем брать не только количество звонков: для SBM полезно будет рассмотреть также в качестве веса суммарную длительность звонков между парой абонентов, а также логарифм данной величины (с прибавлением 1 на случай нулевой длительности). Логарифм может быть полезен для уменьшения влияния экстремально длинных звонков и стабилизации распределения весов. Вершины будут хранить данные о корпусе, где находится абонент, а также описание абонента.

Нахождение оптимального разбиения производится с помощью вызова функции `gt.minimize_blockmodel_dl`, параметрами являются граф, а также словарь `state_args`, в котором передаются: `recs` — указание атрибута ребра для учета взвешенных ребер, `rec_types` — распределение, которое описывает веса ребер, `deg_corr` — применение модифицированного SBM (DC-SBM) для учета наличия хабов в блоках. Разбиение ищется путем минимизации длины описания (description length, количество информации для описания данных с помощью модели), функция быстро находит локально хорошее решение, однако не гарантируется, что оно является наилучшим. В основе лежит MCMC — Markov chain Monte Carlo, этот метод позволяет исследовать пространство разбиений. Алгоритм в `gt.minimize_blockmodel_dl` является агломеративным — он начинает работу с мельчайшего разбиения (каждый узел — отдельный блок), на каждом шаге случайным образом предлагает слить две существующие группы и принимает это сливание только при уменьшении длины описания. Однако данный алгоритм не может разбить блок на два в случае, если это может привести к более выгодному решению.

С помощью вызовов функции `multiflip_mcmc_sweep` можно уточнить решение и найти более оптимальное, однако на это требуется большее время: в нашем случае мы производили 1000 запусков. Это действительно давало стабильное уменьшение длины описания. Данный алгоритм также использует MCMC, но допускает разбиение блока на две части с последующим присоединением к другим блокам, благодаря чему можно найти более оптимальные решения с меньшей длиной описания.

Для проверки стабильности разделения вершин на блоки будем производить 10 запусков SBM и считать ряд метрик по всем возможным парам результатов. Рассматривались метрики ARI (Adjusted Rand Index, значения от -1 до 1, больше — стабильнее) и NMI (Normalized Mutual Information, от 0 до 1, больше

— стабильнее).Из 10 разбиений выбирается лучшее — то есть с минимальной длиной описания.

Рассмотрены несколько вариантов применения SBM.

1. SBM с учетом количества звонков и геометрическим распределением. Предположим, что количество звонков между абонентами будет определять деление на блоки и в паре абонентов количество звонков описывается геометрическим распределением. В результате работы описанной в предыдущем разделе процедуры имеем: $\text{ARI}_{\text{min}} = 0.25, \text{ARI}_{\text{mean}} = 0.47, \text{NMI}_{\text{min}} = 0.39, \text{NMI}_{\text{mean}} = 0.51$. В наилучшем разбиении 12 блоков. Просмотр вершин в блоках показывает, что разбиение не соответствует организационной структуре, не соответствует делению по корпусам. Однако было замечено, что один из блоков полностью соответствует одному из факультетов. Также верно, что одно из подразделений полностью оказалось в одном блоке, однако совместно с другими подразделениями, не имеющими связи с данным.
2. SBM с учетом логарифма длительности звонков и нормальным распределением. Предположим, что логарифм длительности звонков между абонентами будет определять деление на блоки и в паре абонентов логарифм суммарной длительность звонков распределен по нормальному распределению. На самом деле длительность звонка распределена по логнормальному закону, поэтому следовало бы брать распределение, описывающее сумму логнормальных величин, однако из ограниченного выбора распределений наилучший результат по стабильности выдает нормальное распределение. Метрики стабильности: $\text{ARI}_{\text{min}} = 0.17, \text{ARI}_{\text{mean}} = 0.44, \text{NMI}_{\text{min}} = 0.19, \text{NMI}_{\text{mean}} = 0.38$. В наилучшем разбиении 5 блоков, однако соотношения с организационной структурой не наблюдается. Только один блок был выделен таким образом, что состоит из номеров, расположенных в одном корпусе, что дает идею использовать информацию о корпусах.
3. Вложенный SBM. Возможно генерировать вложенное разбиение, при котором узел можно описать как лист в дереве, задаваемом вложенными блоками. Однако применение функции `gt.minimize_nested_blockmodel_dl` с последующим уточнением через `multiflip_mcmc_sweep` с учетом количества звонков и геометрическим распределением дает один уровень вложенности и не представляет интереса.

4. SBM с ограничениями. Используем ограничение (labeled SBM): в один блок не смогут попасть абоненты из разных корпусов. Таким образом мы подсказываем SBM деление на блоки, однако ожидаем, что даже внутри корпуса можно выделить некоторые блоки, например, на основе факультетов. Рассматривается количество звонков, распределение геометрическое. $\mathrm{ARI}_{\mathrm{min}} = 0.65, \mathrm{ARI}_{\mathrm{mean}} = 0.75, \mathrm{NMI}_{\mathrm{min}} = 0.86, \mathrm{NMI}_{\mathrm{mean}} = 0.90$. Значения метрик высокие, однако это вызвано тем, что деление по корпусам уже значительно закрепляет структуру блоков. Три корпуса разделяются по несколько блоков и это разделение связано со структурной организацией.

В рамках предположения, что информация о корпусе может влиять на деление на блоки, было также проверено, как корпуса взаимодействуют между собой. Заметно, что абоненты одного из корпусов часто звонят во многие другие корпуса и также сильно взаимодействуют внутри корпуса. Сделан вывод, что активно телефония используется в четырех корпусах.

В следующем подразделе «Сбор информации об абонентах СГУ» описывается сбор дополнительных номеров телефона с сайта университета. Сбор производится с учетом возникновения возможных ошибок (выполнение нескольких повторов) и реализован с применением многопоточности для ускорения процесса. Собраны номера телефонов подразделений и сотрудников. Подразделения имеют уникальные идентификаторы и образуют иерархическую структуру. Каждый сотрудник также имеет идентификатор, возможно номер(а) телефона(ов) и подразделения, в которых он работает. 560 номеров телефонов указаны как номера сотрудников, 331 номер как телефоны подразделений, с учетом пересечений всего 666 номеров. Из них 471 является городским (имеет префикс 88452). Было выявлено 5 номеров, которые по некоторым признакам могут быть в IP-телефонии, но отсутствуют в базе. 127 номеров из базы IP-телефонии не были найдены на сайте. Городские номера обрезаны до длины 6, остальные номера имеют длину 10.

В заключительном подразделе «Подготовка датасета» описана процедура анонимизации номеров телефонов и формирования датасета. Номера телефонов хэшируются с помощью HMAC-SHA256. Секретный ключ хранится отдельно. Каждый звонок имеет случайно сгенерированный UUID. Время начала и конца представлены в UTC формате, притом часть звонков не имеет времени начала, в этом случае указан null. Оба абонента в каждом звонке имеет один из

типов: ip_sgu (IP-телефония СГУ), other_sgu (не IP-телефония СГУ), external (номера не из СГУ). Статистика по звонкам следующая (пары тип звонящего — тип вызываемого):

- (('ip_sgu', 'external'), 100881);
- (('other_sgu', 'ip_sgu'), 38921);
- (('ip_sgu', 'other_sgu'), 42918);
- (('external', 'ip_sgu'), 94501);
- (('ip_sgu', 'ip_sgu'), 24980);
- (('external', 'external'), 155);
- (('other_sgu', 'external'), 42);
- (('external', 'other_sgu'), 420);
- (('other_sgu', 'other_sgu'), 118).

Последние 4 типа звонков не должны были в целом присутствовать, в датасете звонки такого типа выброшены (тем более их мало).

В итоге датасет содержит 302201 звонков, из них 54106 не содержат времени начала. Каждый звонок содержит 7 атрибутов. Датасет представлен в формате `json`, весит 127 Мб. Вместе с датасетом предлагается список абонентов IP-телефонии СГУ с указанием внешнего хэшированного номера и хэшированного корпуса.

Таким образом, в работе были изучена теория по сложным сетям и существующие датасеты телефонии, после чего была рассмотрена телефонная сеть СГУ. Данные были очищены, а дальше для звонков внутри организации были рассмотрены основные структурные и динамические характеристики с применением современных методов. В результате был получен анонимизированный датасет телефонии СГУ и список абонентов внутри СГУ с указанием пространственной информации, готовые к публикации, которые могут быть полезны для изучения сложных сетей, их моделей, а также для прогнозирования поведения данной сети.

Полученные результаты были представлены на студенческой научной конференции факультета КНиИТ СГУ 23 апреля 2026 года.