МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**

Кафедра дискретной математики и информационных технологий

СОЦИАЛЬНЫЙ ГРАФ И ЕГО ЭМПИРИЧЕСКОЕ ИССЛЕДОВАНИЕ НА ОСНОВЕ МОДИФИЦИРОВАННОЙ МОДЕЛИ БАКЛИ-ОСТГУСА

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента 2 курса 271 группы

направления 09.04.01 – Информатика и вычислительная техника

факультета КНиИТ

Шапошникова Кирилла Сергеевича

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

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

Саратов 2019

ВВЕДЕНИЕ

Изучение сложных объектов, состоящих из большого количества более мелких частей, не сводится к пониманию функционирования этих частей, нужно обобщенное понимание всей структуры, поэтому большую популярность набирает теория сложных сетей. Одной из составляющих частей которой, является веб-граф. Знание об общей структуре веб-графа важно по целому ряду причину, например, из структуры веб-графа мы можем собирать доказательства того что социальный феномен является основополагающим фактором роста Интернета.

Похожим на веб-граф видом графов является социальный граф, который представляет собой сеть пользовательских профилей, хранящих различную информацию: дни рождения, сведения о местах работы, проживания и другую. Изучение данных графов полезно для эффективной генерации рекомендаций и предложений пользователям сети.

Целью данной работы ставится попытка модернизации модели Бакли-Остгуса и эмпирическое исследование социального графа с помощью полученной модели. Актуальность исследования заключается в определении направления для дальнейшего изучения модели Бакли-Остгуса в сфере социальных сетей.

Для достижения цели были сформулированы и решены следующие задачи:

- изучение безмасштабных сетей и лежащих в их основе случайных графов;
- изучение существующих моделей генераторов случайных графов;
- реализация асинхронного генератора графов по модели Боллобаша-Риордана, для последующего ее применения в модернизации метода Бакли-Остгуса;
- построение социального графа по сервису «Твиттер»;
- подготовка классификатора и сбор необходимых характеристик по

социальному графу и графам сгенерированным по рассматриваемым в работе моделям;

- выяснение эмпирическим путем модели, которая лучше всего коррелирует с собранным социальным графом и принятие ее за эталонную в экспериментах по модернизации метода Бакли-Остгуса;
- проведение набора экспериментов с моделью Бакли-Остгуса и анализ полученных результатов.

Методологические основы эмпирического исследования социального графа на основе модифицированной модели Бакли-Остгуса представлены в работах Райгородского [1], Бласиуса [2], Боллобаша [3], Гречникова [4], Штауда [5].

Структура выпускной квалификационной работы включает следующие разделы: введение, пять глав, заключение и список использованных источников.

В первой главе дается понятие безмасштабной сети и описываются ее свойства, а также на основе этого приводится определение случайного графа.

Во второй главе описываются рассматриваемые в работе модели случайных графов: Эредеша-Реньи, Барабаши-Альберт, Боллобаша-Риордана, Бакли-Остгуса, Чунг-Лу.

В третьей главе приводится асинхронный способ реализации модели Боллобаша-Риордана, а также обосновывается выбор технологий для написания данной программы.

В четвертой главе описывается метод анализа моделей случайных графов на основе набора характеристик для выявления моделей лучше коррелирующей с тестовыми данным. Также описывается сбор тестовых данных для анализа, а именно построение социального графа по социальной сети «Твиттер».

В пятой главе приводится эксперимент по модернизации модели Бакли-Остгуса посредством использования динамического параметра начальной предпочтительности вместо статического.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Понятие безмасштабной сети. Безмасштабные сети – это тип сетей, характеризующийся наличием большого количества хабов, также это одни из типов сетей, который обладает законом степенного распределения. Для неориентированных сетей, можно записать закон степенного распределения как:

$$P(q) \propto q^{-\gamma}$$

где γ — показатель степени. $P(q)$ медленно уменьшается при увеличении q, увеличивая вероятность нахождения узла с очень большой степенью. Сети со степенным распределением называются безмасштабными, поскольку законы распределения имеют одинаковую функциональную форму для всех масштабов.

Для большинства реальных сетей постоянная величина γ находится в интервале $2 < \gamma \leq 3$, при этом вероятностное распределение более точно описывается формулой $P(q) \propto (q+q_0)^{-\gamma} \exp(q/q_1)$, где q_0 и q_1 – константы, которые отражают явление насыщения при малых значениях q и обрезания при больших значениях q.

Случайным графом называется случайный элемент из множества $\Omega_t = \{G = (V, E)\}$, которое состоит из графов с $n = n(t)$ вершинами, при этом данная модель характеризуется зависимостью n от t, а также распределением случайного элемента, то есть вероятностями с которыми этот элемент оказывается равным тем или иным конкретным графам G из Ω_t. Основным различием между случайными и безмасштабными сетями является наличие хабов [1].

Модели генерации случайных графов. В разделе описываются модели рассмотренные в работе:

- Эрдеша-Реньи;
- Барабаши-Альберт;
- Боллобаша-Риордана;
- Бакли-Остгуса;

- Чунг-Лу.

В модели Эрдеша-Реньи задается полный граф K_n на множестве вершин V = {1, …, n}, где n – натуральное число. Ребра в K_n нумеруются: e_1, …, e_N, где $N = \binom{n}{2}$. После чего задается некоторое $p \in [0,1]$ и выбираются ребра из e_1, …, e_N согласно схеме Бернулли, то есть каждое ребро независимо от других ребер входит в результирующий граф *G = (V, E)* с вероятностью *p*.

Модель Барабаши-Альберт предполагает, что стартуя с малого m_0 числа вершин, в каждый момент времени добавляется новая вершина с m $(\leq m_0)$ выходящими из нее ребрами, которые соединяют эту новую вершину с m различными вершинами, которые уже имеются в системе. Так как модель имеет предпочтительное присоединение, то предполагается, что вероятность *P* присоединения новой вершины к некоторой вершине *i*, зависит от степени k_i этой вершины, причем:

$$P(k_i) = \frac{k_i}{\sum_j k_j}$$

После *t* шагов модель порождает случайную сеть с $t + m_0$ вершинами и *mt* ребрами.

В модели Чунг-Лу каждому узлу назначается вес, и каждая пара узлов связывается с вероятностью, пропорциональной произведению их весов. В полученном графе каждый узел имеет ожидаемую степень, равную его весу. В качестве веса можно использовать распределение степеней какого-нибудь реального графа. То есть, пусть задано множество вершин $V = \{v_1, …, v_n\}$ и степень каждой вершины d_i, i = 1, …, n. Генерация графа происходит следующим образом [6]:

- Формируется множество L, состоящее из d_i копий v_i для каждого i от 1 до n;
- Задаются случайные паросочетания на множестве L;

- Для вершин u и v из V_s количество ребер в гарфе G, соединяющее их, равно числу паросочетаний между копиями u и v в L.

Метод Боллобаша-Риордана для генерации графа G_m^n, где n – число вершин, а m – параметр, отвечающий за отношение числа ребер к числу вершин, при этом $n, m \in \mathbb{N}$, состоит из двух стадий. На первой стадии строится граф G_1^n посредством индукции, где предполагается, что граф G^{n-1} уже построен и к нему добавляется вершина n с ребром, проводимым по правилу, описываемому формулой [7]:

$$P(i=s)=\begin{cases}\dfrac{d_{G_1^{n-1}}(s)}{2n-1} & \text{если } 1\le s\le n-1,\\ \dfrac{1}{2n-1} & \text{если } s=n\end{cases}$$

После того как был получен граф G_1^n, алгоритм переходит ко второму этапу, на котором его вершины «схлопываются». Для этого вершины объединяются в группы по m вершин и каждой такой группе даются собственные обозначения v_i, где $i \in \{1,\dots,n-1\}$, например вершина v_1 = {*1, ..., m*}, а вершина v_2={*m+1,...,2m*}. В каждой вершине v_i строится столько петель сколько ребер было между вершинами соответствующей группы исходного графа. Аналогично строятся ребра между вершинами *v*, ребро проводится в случае, если есть ребро между соответствующими группами исходного графа. На выходе из графа G_1^{mn} получается граф G_m^n [3].

Модель Бакли-Остгуса похожа на модель Боллобаша-Рииордана, но вводит дополнительный положительный коэффициент *a,* не зависящий от степени и называемый начальной привлекательностью или предпочтительностью вершины. При этом алгоритм добавления новых вершин и ребер в генерируемый граф не изменяется, лишь вероятности проведения ребра в новую вершину начинают вычисляться с учетом коэффициента *a* [4]:

$$P(i=s)=\begin{cases}\dfrac{d_{H_{a,1}^{n-1}}s+a-1}{(a+1)n-1} & \text{если } 1\le s\le n-1,\\ \dfrac{a}{(a+1)n-1} & \text{если } s=n\end{cases}$$

Реализация асинхронного генератор. В данном разделе описывается способ распараллеливания модели Боллобаша-Риордана.

Моделирование графа состоит из двух этапов, поэтому они рассматривались независимо друг от друга. На первом этапе было решено сохранить добавление новых вершин в один поток, а расчет вероятностей распараллелить, поскольку из-за добавления новой вершины на каждом шаге, это действие нужно повторять и с ростом графа объем вычислений постепенно увеличивается. Массив, содержащий степени вершин, разделяется на количество процессоров в системе, соответственно вероятности вычисляются в отдельных потоках, после чего рассчитанные части соединяются в один массив, сохраняя при этом исходный порядок.

Второй этап метода Боллобаша-Риордана хорошо поддается распараллеливанию поскольку исходный граф G_1^{mn} уже построен, поэтому здесь не требуется введение в код дополнительных средств синхронизации, то есть можно сразу разбить последовательность вершин G_1^{mn} на отрезки, которые будут обрабатываться параллельно, в результирующем графе порядок следования вершин ни на что не влияет, поэтому сортировка, как на предыдущем шаге, не требуется.

Описанный генератор был реализован на языке программирования Go, поскольку он реализует механизм горутин и каналов, позволяющих удобно работать с асинхронным кодом. Сгенерированные графы хранятся в формате JSON в NoSql базе данных MongoDB.

Классификация моделей. Раздел содержит описание методики анализа моделей случайных графов для выявления модели лучше коррелирующей с тестовыми данными.

Методика строится на сборе и классификации характеристик по каждому графу, сгенерированному по какой-то модели, и также по тестовому графу. Собираемые характеристики делятся на две категории: характеристики с одним значением и характеристики с размерностью, равной количеству узлов графа. К первой категории относятся: количество узлов графа (n), количество ребер (m), диаметр, а также эффективный диаметр. Поскольку предполагается, что распределение степеней в графах должно подчиняться степенному закону, то в данную категорию также входит и экспонента (γ) степенного закона

$$P(q) \propto q^{-\gamma}$$

Во вторую категорию входят характеристики, размерность которых зависит от числа узлов графа – это, например, всевозможные центральности. Список таких характеристик следующий:

- степенное распределение;
- центральность по посредничеству;
- центральность по близости;
- центральность по Кацу;
- пейджранк центральность.

Так как центральности имеют различную размерность для графов с разным числом вершин, то производится «свертка» векторов центральностей, то есть для каждого вектора вычисляются параметры;

- среднее значение,
- медиану,
- первый и третий квартили,
- максимальное и минимальное значение
- дисперсия
- коэффициент эксцесса
- коэффициент ассиметрии.

Для сборки тестового графа, необходимого для проведения исследования, была написана программа-crawler на языке Go, основанная на публичном API

«Твиттер». Собранный граф был сохранен как список смежности в MongoDB базу данных в JSON формате для удобного обращения к любой вершине и более быстрой загрузки всего графа. Собранный граф подчиняется степенному распределению и имеет характеристики перечисленные в таблице 1.

Таблица 1 – Параметры графа пользователей участка сети Твиттер.

Характеристика	Значение
Узлы, ребра	13544931, 34885363
Изолированные вершины	0
Петли	0
Плотность	0.00000038
Кластерный коэффициент	0.084986
Минимальная, максимальная, средняя степени	1, 15476, 5.151058
Коэффициент ассортативности	-0.236831
Количество компонент связности	2
Размер гигантской компоненты связности	13544800 (100%)
Экспонента степенного закона	2.255

Классификация моделей происходит с помощью библиотеки sklearn в связи с наличием реализации SVM классификатора. Работа SVM похожа на Линейный Регрессионный классификатор, где между двумя классами проводится граница линейного решения. Различие же между этими двумя классификаторами заключается в том, что в SVM граница проводится так, чтобы расстояние между точками около нее было максимальным. Точки же на границах называются опорными векторами. Вычисленные характеристики подаются на вход классификатора попарно, поскольку в случае наличия двух классов SVM работает точнее. Также важным моментом является то что у графов брались их гигантские компоненты связности, поскольку центральности могут быть вычислены только по связным графам.

В качестве результата классификатор может попытаться явно предсказать класс, к которому относится тестовый набор данных, но в нашем случае

удобнее смотреть на вероятность, с которой каждый класс обучающего набора относится к тестовому. Для каждого класса было взято по 10 векторов характеристик. Тестовый граф разделялся на подграфы, чтобы можно было получить необходимый набор данных для исследования. Результаты классификации моделей Эрдеша-Реньи, Барабаши-Альберт и Чунг-Лу представлены в таблице 2.

Таблица 2 – Результаты попарного сравнения различных моделей.

Пара моделей	Вероятность соответствия тестовому классу
Эрдеша-Реньи – Чунг Лу	0.29013239 – 0.70986761
Барабаши-Альберт – Чунг Лу	0.2379145 – 0.7620855
Барабаши-Альберт – Эрдеша-Реньи	0.64091414 – 0.35908586

Эксперименты с моделью Бакли-Остгуса. В разделе описывается постановка и проведение эксперимента по модернизации модели Бакли-Остгуса.

В модели Бакли-Остгуса используется статический коэффициент начальной предпочтительности, поэтому было решено провести эксперимент по использованию динамического параметра, изменяющегося в интервале от l до r с шагом s, при этом параметры l, r, s становятся входными параметрами модифицированной модели. Изменение коэффициента a в выбранном интервале производится после добавления каждой новой вершины на первом этапе модели Бакли-Остгуса.

Перед проведением эксперимента была выбрана эталонная модель с которой будет сравниваться модифицированная модель, такой моделью оказалась модель Чунг-Лу, для выбора эталона использовалась программа классификации моделей описанная ранее.

Серия экспериментов с моделью Бакли-Остгуса включала изменение параметра a с разным шагом и в разных интервалах, а также варианты с направлением изменения коэффициента, то есть его уменьшения или

увеличения. В качестве реализации модели использовался модифицированный генератор модели Боллобаша-Риордана, ранее описанный в работе.

В результате эксперимента было выяснено, что лучшие результаты были показаны при уменьшении коэффициента a в интервале 1, …, 0.47 с шагом 0.01, рисунок 1. При этом граница 0.47 была выбрана заранее, как лучший коэффициент начальной предпочтительности исходной модели Бакли-Остгуса при сравнении ее графов с тестовым графом собранным по сети «Твиттер».

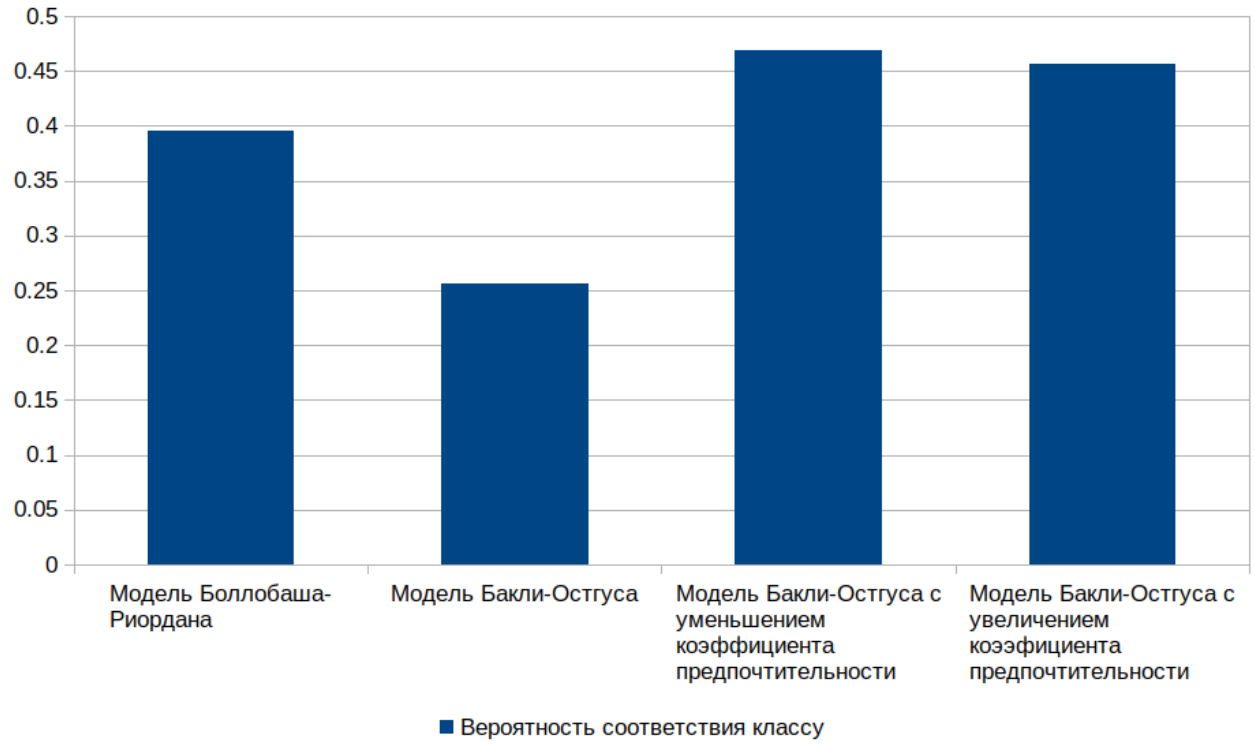

Рисунок 1 – Результаты сравнения моделей основанных на методе Боллобаша-Риордана с Чунг-Лу.

ЗАКЛЮЧЕНИЕ

В результате проделанной работы было рассмотрено понятие безмасштабных сетей и лежащих в их основе случайных графов. Также были изучены модели построения таких сетей и на основе модели Боллобаша-Риордана был разработан генератор, который поддерживает асинхронный режим работы на мультипроцессорной системе.

Разработанный генератор позволяет экономить время при генерации графов, содержащих большое количество узлов. Проведенный эксперимент по генерации графов различной размерности позволяет сделать вывод, что при использовании четырехядерной машины распараллеливание алгоритма значительно уменьшает время построения графа.

Данный генератор также применялся для генер=ации графов по модели Бакли-Остгуса, модификация которой и выяснение применимости на реальных социальных графах, являлась основной целью работы. Для проведения такого эксперимента и получения реального социального графа, была разработана программа-crawler, которая предназначена для обхода социальной сети Твиттер на заданную глубину, начиная с указанного пользователя. Также программа позволяет сохранить результаты работы в базу данных для дальнейшего анализа.

После получения тестового социального графа, были также сгенерированы графы и собраны их характеристики по моделям Эредеша-Реньи, Барабаши-Альберт, Боллобаша-Риордана и Чунг-Лу. Для сравнения их характеристик, была разработана программа на основе SVM классификатора, которая возвращает вероятность принадлежности тестовых данных переданным на вход моделям на основе вычисленного вектора характеристик каждого графа.

В качестве модификации модели Бакли-Остгуса было выбрано динамическое изменение параметра начальной предпочтительности a в заданных пределах и с определенным шагом. Результаты эксперимента показали состоятельность такой модификации, поскольку результаты,

показываемые модифицированной модели оказались лучше, чем у исходной модели Бакли-Остгуса и модели Боллобаша-Риордана, которая является частным случаем метода Бакли-Остгуса, при сравнении с моделью Чунг-Лу, выбранной в качестве эталонной модели. Кроме того, данная модификация позволяет более гибко подстраивать модель для реальных графов, благодаря наличию дополнительных входных параметров.

По результатам работы были опубликованы статьи [8, 9, 10].

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Райгородский А. М., Модели интернета: Учебное пособие – Интеллект, 2013. – 64 с.

2 Bläsius T., Friedrich T., Katzmann M., Krohmer A., Striebel J. (2018) Towards a Systematic Evaluation of Generative Network Models. In: Bonato A., Prałat P., Raigorodskii A. (eds) Algorithms and Models for the Web Graph. WAW 2018. Lecture Notes in Computer Science, vol 10836. Springer, Cham

3 B. Bollobas. Random Graphs, Second Edition. – Cambridge Univ. Press, 2001.

4 E. A. Grechnikov, G. G. Gusev, L. A. Ostroumova, Yu L. Pritykin, A. M. Raigorodskii, P. Serdyukov, D. V. Vinogradov, and M. E. Zhukovskiy. Empirical validation of the buckley–osthus model for the web host graph. In The 21st ACM Conference on Information and Knowledge Management, pages 1577–1581, 2012.

5 Staudt, C. L., Sazonovs, A., Meyerhenke, H,: NetworKit: a tool suite for large-scale complex network analysis. Netw. Sci. 4(4), 508-530 (2016)

6 Chung, F., Lu, L.: Connected components in random graphs with given expected degree sequences. Ann. Comb. 6(2), 125–145 (2002)

7 Райгородский А. М., Модели случайных графов – М.: МЦНМО, 2011. – 136 с.

8 Шапошников К. С., Сагаева И. Д., Сидоров С. П., Компьютерные науки и информационные технологии: Материалы международной научной конференции // Реализация асинхронного генератора графов методом Боллобаша-Риордана. – Саратов: Издат. Центр «Наука», 2018. – 464 с.

9 Шапошников К. С., Сагаева И. Д., Сидоров С. П., Информационные технологии и математическое моделирование (ИТММ-2019): Материалы XVIII Международной конференции имени А.Ф. Терпугова (26−30 июня 2019 г.) // Генерация сложных сетевых структур на основе оптимизированной модели с предпочтительным присоединением. – Томск: Издат. НТЛ, 2019. – Часть 2. – с. 75-79.

10 Шапошников К. С., Сагаева И. Д., Сидоров С. П., Информационные технологии и математическое моделирование (ИТММ-2019): Материалы XVIII Международной конференции имени А.Ф. Терпугова (26−30 июня 2019 г.) // Анализ сложных сетевых структур на примере социальной сети «Twitter». – Томск: Издат. НТЛ, 2019. – Часть 2. – с. 71-74.