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

СЕТЕВЫЕ МОДЕЛИ НА ОСНОВЕ МЕХАНИЗМОВ ПРЕДПОЧТИТЕЛЬНОГО ПРИСОЕДИНЕНИЯ, ПРИСПОСОБЛЕНИЯ И ТРОЙСТВЕННОГО ЗАМЫКАНИЯ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента 2 курса 271 группы
направления 09.04.01 — Информатика и вычислительная техника
факультета КНиИТ
Ступака Александра Васильевича

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

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

Саратов 2023

ВВЕДЕНИЕ

Сетевая модель обладает свойствами, аналогичными реальным системам. Сети считаются мощным инструментом для представления моделей связей между частями систем, таких как Интернет, электросети, продовольственные и социальные сети. Огромное количество систем во многих отраслях науки можно смоделировать в виде больших разреженных графов, обладающих многими геометрическими свойствами. Одна из важнейших задач, которую помогают решить сетевые модели — это анализ социальных данных, который стремительно набирает популярность во всем мире благодаря появлению в 1990-х годах онлайновых сервисов социальных сетей.

Таким образом, модели генераторов случайных графов и их применение для эмпирических исследований социальных сетей становятся особенно актуальными в современном мире.

Целью выпускной квалификационной работы является моделирование и исследование реальных динамических сетей на основе пользовательских данных из социальных сетей.

Из данной цели вытекают следующие задачи:

— рассмотреть модель механизма предпочтительного присоединения и приспособления;
— рассмотреть модель механизма тройственного замыкания;
— изучить такие характеристики реальных графов, как усредненная средняя степень соседей вершин, степенной закон, коэффициент кластеризации, центральность «по близости» и «посредничеству»;
— сгенерировать динамические модели искусственных сетей, основанных на характеристиках реальных графов;
— сравнить характеристики реальных и искусственных сетей.

Структура и объем работы.

Для решения поставленных задач выполнена выпускная квалификационная работа, которая включает в себя введение, 9 основных глав, заключение, список использованных источников из 25 наименований и 6 приложений. Первая глава имеет название «Предпочтительное присоединение в сетях». Вторая глава имеет название «Модель с приспособленностью: сеть Бьянкони-Барабаши». Третья глава имеет название «Тройственное замыкание как основной механизм создания сообществ в сложных сетях». Четвер-

тая глава имеет название «Вычисление параметров моделей искусственных сетей». Пятая глава имеет название «Описание характеристик реальных и искусственных сетей». Шестая глава имеет название «Инструменты разработки». Седьмая глава имеет название «Реализация программ построения искусственных сетей». Восьмая глава имеет название «Реализация программ вычисления характеристик». Девятая глава имеет название «Вычисление и сравнение характеристик для реальных и искусственных сетей». Работа изложена на 68 страницах, содержит 23 формулы, 19 рисунков и 1 таблицу.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе описан механизм построения искусственной сети, на котором основана модель предпочтительного присоединения. Барабаши и Альберт в работе [1] предложили два основных механизма, которые пытаются лучше охарактеризовать реальную сеть: рост системы, а именно добавление новых узлов и предпочтительное присоединение, когда новый узел преимущественно соединяется с наиболее подключенными узлами, уже находящимися в сети. Например, сеть Интернет расширяется с добавлением новых документов, которые ссылаются на старые или хорошо известные сайты. Вероятность того, что новый узел соединится с узлом с k ссылками, пропорциональна k, независимо от географического расстояния.

Несмотря на то, что модель Барабаши-Альберта объясняет основные особенности большей части сетей реального мира (степенной характер распределения степеней, динамику роста степеней отдельных узлов), строго фиксированное значение показателя $\gamma = 3$ является существенным недостатком модели. Анализ сетей реального мира, приведенный в статье [2], показал, что значение показателя степени $1 \leq \gamma \leq 3$. Очевидно, что эта особенность связана с присутствием дополнительных механизмов роста сетей.

Во второй главе описывается модель приспособленности Бьянкони-Барабаши. Оригинальная модель Барабаши-Альберта создает сеть, которая основана на законе предпочтительного присоединения, где узлы с большей степенью в сети становятся привилегированными для создания новых связей. Но эта модель не учитывает конкурентоспособность, например, способность более новых узлов приобретать соседей.

Чтобы смоделировать такую ситуацию, можно включить внутреннюю характеристику в каждом узле, называемую приспособленностью. В социальных сетях приспособленность будет представлять собой признак того, что человек становится более популярным благодаря какому-то своему качеству. Чем выше приспособленность, тем выше вероятность привлечения новых ребер. Приспособленность можно определить как способность привлекать новые ссылки — количественный показатель способности узла оставаться впереди своих конкурентов.

Эта характеристика наблюдалась в сетях Бьянкони и Барабаши, описанных в статье [3]. Позже они предложили альтернативную модель, вклю-

чающую коэффициент приспособленности η_i каждого узла i. Алгоритм практически аналогичен сети Барабаши-Альберта, но каждый узел подключается к существующему узлу в сети с вероятностью, которая, помимо зависимости от степени k, также пропорциональна приспособленности η_i. Выбор значений $\eta_i \in [0, 1]$ может определяться равномерным распределением $\rho(\eta_i)$.

В третьей главе описывается модель тройственного замыкания — механизма, который является сильным механизмом-кандидатом для создания связей в графах, особенно в социальных сетях. При предпочтительном присоединении именно степень узла определяет вероятность соединения, подразумевая, что каждый новый узел знает эту информацию обо всех других узлах, что нереально. Вместо этого тройственное замыкание вызывает эффективное предпочтительное присоединение: соединение с соседом A узла соответствует выбору A с вероятностью, возрастающей со степенью k_A этого узла, в соответствии с линейным предпочтительным присоединением, описанным в статьях [4-5]. Этот принцип лежит в основе нескольких генеративных сетевых моделей, описанных в статье [6], все из которых дают графики с распределениями степеней с высокими коэффициентами кластеризации.

В четвертой главе описаны параметры моделей искусственных сетей. В алгоритмах, основанных на механизмах предпочтительного присоединения, приспособления и тройственного замыкания, присутствуют параметры n и m, обозначающие конечное количество вершин в сети и количество присоединяемых ребер на одном шаге итерации соответственно. Для того, чтобы искусственная сеть была наиболее приближена по основным характеристикам к реальной сети, было принято взять за n количество вершин в реальной сети, которая будет сравниваться по характеристикам с искусственной сетью, а также вычислить такое значение m, которое максимально приблизит количество связей $|E|$ к значениям реального графа.

В алгоритмах, основанных на механизме тройственного замыкания, присутствует такой параметр, как p — вероятность присоединения к соседней вершине. Его значение зависит от коэффициента кластеризации сети и варьируется в диапазоне от 0 до 1.

Оптимальное значение параметра p высчитывается экспериментальным путем в ходе сравнения реальных и искусственных сетей по такой характеристике, как коэффициент кластеризации.

В пятой главе описываются характеристики реальных и искусственных сетей. Для исследования реальных сетей была выбрана такая величина, как средняя степень соседей. Данная величина используется при анализе степенных корреляций и ассортативности в сложных сетях. Ассортативность, о которой подробно рассказано в статье [7], отражает предпочтение узлов присоединяться к схожим узлам, в частности к узлам с близкой по значению степенью, как описано в статьях [8-10].

Следующая важная характеристика сетей — распределение степеней узлов. Распределение степеней $P(k)$ графа определяется как доля узлов, имеющих степень k. В сложных или социальных сетях работают специфические законы распределения степеней узлов. В частности, в сетях большинство узлов имеют низкую степень, но в то же время встречаются узлы с высокой степенью. В природе встречается очень много случаев, когда законы распределения — степенные. Авторы статьи [11] показали, что Распределение степеней вершин реальных графов социальных сетей хорошо аппроксимируется степенной зависимостью, так что основным направлением моделирования является построение случайных графов с таким распределением.

Кластеризация — это локальная характеристика сети. Она характеризует степень взаимодействия между собой ближайших соседей данного узла. Коэффициент кластеризации данного узла есть вероятность того, что два ближайших соседа этого узла сами есть ближайшие соседи. Он может быть усреднен для любой части сети или для сети в целом, становясь ее интегральной характеристикой. Данный параметр характеризует степень связности графа.

Для исследования сетей были выбраны такие величины, как центральность «по близости» и «посредничеству», описанные в статье [12]. Самой понятной является концепция, в основе которой лежит предположение, что вершина с большим значением степени является самой важной. Иной взгляд на центральность основывается на частоте прохождения через вершину кратчайших путей между всеми парами вершин и называется посредничеством. Считается, что вершина с наибольшим значением центральности «по посредничеству» наибольшим образом влияет на группы остальных вершин, так как является основной при транзите информации через нее.

Следующая концепция определения центральности вершины «по близо-

сти» также базируется на идее контроля отношений между вершинами графа. Центральную позицию занимает вершина, наиболее независимая от остальных вершин при передаче информации. В таком случае центральность «по близости» измеряется путем суммирования расстояний от вершины до всех остальных. Центральность «по близости» выражает, насколько близко узел расположен к остальным узлам сети. Центральность «по близости» является показателем того, насколько быстро распространяется информация в сети от одного участника к остальным, то есть насколько близок рассматриваемый участник ко всем остальным участникам сети.

Наиболее распространенный подход для рассмотрения понятия центральности графа в целом — рассматривать центральность графа как степень отличия значения меры центральности вершины с наибольшим значением от значений для других вершин.

В шестой главе описаны инструменты разработки. Программы для моделирования и исследования реальных сетей были написаны на языке Python. Входные данные для анализа сложных сетей были взяты из онлайн-репозиториев. Для моделирования реальных и построения искусственных сетей была использована библиотека NetworkX, написанная на языке программирования Python. Как упоминает публикация [13], NetworkX — это пакет Python для создания, манипулирования и изучения структуры, динамики и функций сложных сетей. С NetworkX можно загружать и хранить сети в стандартных и нестандартных форматах данных, генерировать множество типов случайных и классических сетей, анализировать структуру сети, строить сетевые модели, проектировать новые сетевые алгоритмы, рисовать сети и многое другое.

В седьмой главе описывается реализация программ построения искусственных сетей на основе алгоритмов предпочтительного присоединения, приспособления и тройственного замыкания. Искусственная сеть в программе строится с помощью записи в файл набора ребер, которые образуются после прохождения алгоритма.

В восьмой главе описывается реализация программ вычисления характеристик реальных и искусственных сетей.

Чтобы измерить среднюю степень соседей, были взяты определенные небольшие наборы вершин (3 набора по 100 вершин). На каждом временном

шаге для каждой вершины набора считалась средняя степень ее соседей, которая затем усреднялась по всему набору. Итоговое значение — усредненная средняя степень соседей.

Чтобы построить график распределения степеней узлов, в программе ищется количество вершин каждой степени вплоть до максимальной. Для более удобного отображения значения массивов для графика сразу прологарифмированы.

Чтобы подсчитать коэффициент кластеризации, нужно для каждого узла сети подсчитать локальный коэффициент, а затем найти среднее значение по всему графу.

Для вычисления центральности «по близости» для всех узлов сети используется функция closeness_centrality из библиотеки NetworkX. Она возвращает массив пар вида «вершина – центральность вершины». Далее вычисляются общие значения для этих центральностей — максимум, минимум, среднее значение, стандартное отклонение, коэффициент вариации. Помимо этого, вычисляется главная характеристика по всей сети — центральность графа.

Для вычисления центральности «по посредничеству» для всех узлов сети используется функция betweenness_centrality из библиотеки NetworkX. Она также возвращает массив пар вида «вершина – центральность вершины». Алгоритм и необходимые значения вычисляются также, как и для центральности «по близости».

В девятой главе описаны результаты вычисления и сравнения характеристик реальных и искусственных сетей.

Модель Барабаши-Альберта предполагает, что большинство узлов будут иметь низкую степень, а оставшиеся немногочисленные вершины будут иметь высокую степень, как это наблюдалось во множестве реальных сетей. Эмпирически было установлено, что модель предпочтительного присоединения генерирует сеть с распределением степеней $P(k) \sim k^{\gamma}$ при $\gamma = -3.0$, независимо от значения параметра m.

Что касается коэффициента кластеризации, то в реальных сетях, как правило, его значение оказывается больше, чем в сетях, построенных на основе алгоритма предпочтительного присоединения, что и показывают построенные в работе диаграммы. Поэтому для построения более адекватных моделей

требуется модификация процедуры предпочтительного присоединения, позволяющая увеличивать коэффициент кластеризации, не изменяя при этом распределения степени вершин. Для приближения к реальным значениям была разработана модель тройственного замыкания. Процесс регулирования коэффициента кластеризации основывается на изменении количества образуемых в процессе генерации «треугольников».

У данной модели три главных входных параметра:

— количество вершин n;

— количество присоединяемых ребер на одном шаге итерации m;

— вероятность присоединения к соседней вершине p.

Первые два параметра высчитываются на основе данных реальной сети, а оптимальное значение вероятности p находится эмпирически.

В ходе экспериментальной проверки было обнаружено наиболее оптимальное значение параметра модели $p = 0,005$.

После серии экспериментов было выявлено, что модель построения искусственной сети, основанная на алгоритме тройственного замыкания, более точно описывает реальную модель, чем механизм предпочтительного присоединения.

Для сравнения значений центральности «по близости» и «посредничеству» реальной и искусственных сетей были взяты 3 сети:

— реальная сеть, основанная на данных социальной сети Facebook;

— искусственная сеть, построенная по алгоритму предпочтительного присоединения Барабаши-Альберта;

— искусственная сеть, построенная по алгоритму Бьянкони-Барабаши.

Модель Бьянкони-Барабаши показала значения, более приближенные к реальным, в отличие от модели Барабаши-Альберта.

После серии экспериментов было выявлено, что модель построения искусственной сети, основанная на алгоритме Бьянкони-Барабаши ввиду дополнительной третьей концепции приспособленности более точно описывает реальную модель, чем механизм предпочтительного присоединения Барабаши-Альберта, с точки зрения центральности графа.

ЗАКЛЮЧЕНИЕ

В работе было выполнено следующее:

— рассмотрены модели механизма предпочтительного присоединения, приспособления и тройственного замыкания;

— изучены такие характеристики реальных графов, как усредненная средняя степень соседей вершин, степенной закон, коэффициент кластеризации, центральность «по близости» и «посредничеству»;

— сгенерированы динамические модели искусственных сетей, основанных на характеристиках реальных графов;

— получены результаты сравнения характеристик реальных и искусственных сетей.

Результаты эмпирического исследования позволяют сделать вывод, что реальные сети, основанные на данных социальной сети Facebook и сайта Stack Overflow, достаточно хорошо моделируются моделями предпочтительного присоединения, приспособления и тройственного замыкания с точки зрения распределения степеней, центральности графа и кластеризации.

Основные источники информации:

1 Barabashi, A.L. Emergence of scaling in random networks / A.L. Barabashi, R.R. Albert // Department of Physics. — 1999. — Vol. 286. — Pp. 509-512.

2 Albert, R. Statistical mechanics of complex networks / R. Albert, A.-L. Barabashi // Reviews of Modern Physics. — 2002. — Vol. 74. — Pp. 43-97.

3 Bianconi, G. Competition and multiscaling in evolving networks / G. Bianconi, A.-L. Barabashi // EPL (Europhysics Letters). — 2001. — Vol. 54.

4 Шапошников, К.С. Генерация сложных сетевых структур на основе оптимизированной модели с предпочтительным присоединением / 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.). — 2019. — № 2. — С. 75-79.

5 Shaposnikov, K. Random graph models and their application to twitter network analysis / Shaposhnikov K., Sagaeva I., Grigoriev A., Faizliev A., Vlasov A. // Series: Atlantis Highlights in Computer Sciences Proceedings of the Fourth Workshop on Computer Modelling in Decision Making (CMDM

2019). — 2019.

6 Aynaud, T. Multilevel local optimization of modularity / T. Aynaud, V.D. Blondel, J.-L. Guillaume, R. Lambiotte // John Wiley and Sons. — 2013. — Pp. 315-345.

7 Fisher, David N. The Perceived Assortativity of Social Networks: Methodological Problems and Solutions / David N. Fisher, Matthew J. Silk, Daniel W. Franks // Trends in Social Network Analysis: Information Propogation, User Behavior Modeling, Forecasting, and Vulnerability Assessment / Ed. by Rokia Missaoui, Talel Abdessalem, Matthieu Latapy. — Cham: Springer International Publishing, 2017. — Pp. 1-19.

8 Barabashi, A.L. Network science / A.L. Barabashi // Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. — 2013. — Vol. 371, no. 1987.

9 Grigoriev, A. Average nearest neighbor degree and its distribution in social networks / Grigoriev A., Sidorov S., Mironov S., Malinskii I. // Digital Transformation and Global Society. DTGS 2021. Communications in Computer and Information Science. — 2022. — Vol. 1503. — Pp. 36-50.

10 Yao, D. Average nearest neighbor degrees in scale-free networks / Yao D., van der Hoorn P., Litvak N. // Internet mathematics. — 2018. — Vol. 1. — Pp. 1-38.

11 Кислицын, А.А. Модель эволюции распределений степеней вершин графов социальных сетей / А.А. Кислицын, Ю.Н. Орлов // Математическое моделирование. — 2021. — Т. 33, № 9. — С. 3-21.

12 Bonchi, F. Centrality measures on big graphs: exact, approximated and distributed algorithms / F. Bonchi, G. De Francisci, M. Riondato // Proc. 25th International Conference Companion on World Wide Web. — 2016. — Pp. 1017-1020.

13 Hagberh, Aric A. Exploring network structure, dynamics, and function using networkx // Proceedings of the 7th Python in Science Conference / Ed. by Gael Varoquaux, Travis Vaught, Jarrod Millman. — Pasadena, CA USA: 2008. — Pp. 11-15.