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

Кафедра дискретной математики и информационных технологий

ГИБРИДНЫЙ АНАЛИЗ НАУЧНЫХ СЕТЕЙ: ТОПОЛОГИЧЕСКИЕ МЕТРИКИ СВЯЗНОСТИ И СЕМАНТИЧЕСКОЕ ПРОФИЛИРОВАНИЕ АВТОРОВ НА ОСНОВЕ ВЕКТОРНЫХ МОДЕЛЕЙ ТЕКСТОВЫХ КОРПУСОВ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента 2 курса 271 группы
направления 09.04.01 — Информатика и вычислительная техника
факультета КНиИТ
Крылова Романа Алексеевича

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

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

Саратов 2026

ВВЕДЕНИЕ

Актуальность темы. Сетевые структуры представляют собой неотъемлемый компонент исторического развития человечества, выступая фундаментальным организационным принципом, на основе которого исторически формировались и продолжают эволюционировать ключевые аспекты социальной действительности. Указанные структуры оказывают непосредственное влияние не только на характер межличностных взаимодействий и механизмы формирования общественного мнения, но и играют определяющую роль в процессах экономического развития, обеспечения политической стабильности, а также в поддержании культурного многообразия. В связи с этим глубокое понимание роли и механизмов функционирования сетевых образований выступает необходимым условием для комплексного осмысления структурных закономерностей, лежащих в основе организации современного мира. Особый интерес представляет научная коллаборация как частный случай сложной сети, где узлами выступают исследователи, а ребра — отношения соавторства. Изучение поведенческих паттернов групповых образований в таких сетях, а также выявление устойчивых закономерностей, характеризующих внутренние процессы взаимодействия и обмена информацией между элементами данных структур, является актуальной научно-практической задачей.

Цель исследования — комплексное изучение поведенческих паттернов групп в рамках топологии сложных сетей соавторства, а также выявление устойчивых закономерностей, характеризующих внутренние процессы взаимодействия и обмена информацией между элементами данных структур.

Для последовательной реализации поставленной цели исследовательский процесс структурирован в четыре взаимосвязанных главы:

1. **Теоретико-методологический обзор**, предполагающий систематический поиск и критический анализ научной литературы, посвященной вопросам конструирования и характеризации сложных сетей, а также подбор и первичный анализ эмпирического набора данных с целью оценки его репрезентативности и пригодности для последующего моделирования;

2. **Топологический анализ**, направленный на выявление узлов централизации исследуемой структуры и проведение оценки связности графовых компонент, входящих в состав сети;
3. **Эмпирический анализ**, включающий проведение всестороннего контент-анализа отдельных узловых элементов сети, что позволило детализировать их структурные и функциональные особенности;
4. **Векторный анализ**, на котором осуществляется выявление и систематизация закономерностей, определяющих поведение ключевых узлов в составе групповых кластеров, а также оценка степени их влияния на динамику данных групп.

Для обеспечения методологической последовательности и достижения заявленной цели были сформулированы следующие исследовательские задачи:

1. Осуществить систематический поиск и рецензирование профильной научной литературы, формирующей теоретико-методологическую базу исследования;
2. Выявить и отобрать релевантный эмпирический набор данных, удовлетворяющий критериям полноты и достоверности;
3. Сконструировать модель сложной сети на основе отобранного массива данных с применением соответствующих алгоритмов графового моделирования;
4. Провести структурный и метрический анализ построенной сетевой модели для определения ее базовых характеристик;
5. Реализовать программно процедуру контент-анализа, направленную на глубокую интерпретацию содержательных атрибутов сетевых узлов;
6. Выявить специфические механизмы влияния ключевых узлов на динамику соответствующих групповых образований, а также оценить их воздействие на глобальную архитектуру исследуемой сети.

Краткая характеристика материалов исследования. В качестве эмпирической базы использован открытый репозиторий DBLP (Digital Bibliography & Library Project), содержащий детализированные сведения об участии исследователей в научных конференциях. Репозиторий охватывает один миллион записей, ретроспективу публикационной активности за 78 лет и включает информацию о более чем 950 тысячах уникальных авторов. Данные

организованы в строго формализованном формате, обеспечивающем высокую степень машиночитаемости и упрощающем процедуру автоматизированного извлечения графовых структур. Для обработки текстовых данных (названий и аннотаций публикаций) задействован инструментарий языка Python, в частности библиотека Natural Language Toolkit (NLTK).

Структура магистерской работы. Работа состоит из введения, четырех глав, заключения, списка использованных источников и четырех приложений.

— **Глава 1** содержит теоретико-методологический обзор: систематизацию ключевых понятий теории графов и сетевого анализа, описание генеративных моделей сетевых структур, а также обоснование выбора и предварительный анализ эмпирической базы данных DBLP.

— **Глава 2** посвящена топологическому анализу: исследованию точек централизации в сети публикаций и соавторства, а также анализу структуры и динамики компонент связности соавторской сети.

— **Глава 3** описывает эмпирический анализ: методологию контент-анализа, процедуры предобработки данных, алгоритм формирования словаря ключевых единиц и эмпирический анализ частотных характеристик лексических единиц.

— **Глава 4** представляет векторный анализ и интерпретацию результатов: теоретическое обоснование векторной модели представления текстовых корпусов, обзор метрик сравнения текстовых векторов, обоснование выбора метрики, а также анализ лексической близости на примере узла с максимальной степенью связности.

Научная новизна работы заключается в комбинированном применении методов топологического сетевого анализа и векторных моделей текстов для исследования научной коллаборации. В частности, в работе предложена методика количественной оценки семантической близости соавторов через косинусное сходство их лексических профилей, построенных на основе частотного распределения терминов в названиях и аннотациях публикаций, и продемонстрирована корреляция между интенсивностью кооперационного взаимодействия и степенью лексической согласованности авторов.

Научная значимость работы состоит в выявлении устойчивых закономерностей фрагментации сети соавторства, подтверждении баланса между процессами появления новых изолированных компонент и интеграции существующих узлов в крупные кластеры, а также в демонстрации того, что совместная публикационная деятельность способствует конвергенции терминологического аппарата соавторов. Полученные результаты могут быть использованы для прогнозирования формирования новых научных коллабораций, анализа эволюции исследовательских областей и разработки рекомендательных систем в академической среде.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В первом разделе работы проведен теоретико-методологический обзор, формирующий концептуальную базу исследования. Сложные сети определены как математические модели систем, состоящих из множества дискретных компонентов (узлов или вершин) и связей (ребер или дуг) между ними. Формализация таких структур опирается на аппарат теории графов, однако ключевым отличием современных исследований является фокус на системах с неочевидной, нетривиальной топологией, которая не может быть сведена к регулярным решеткам или полностью случайным графам.

Систематизированы ключевые определения, используемые в ходе исследования. Граф (сеть) определен как математическая структура $G = (V, E)$, где V — множество вершин, E — множество ребер. Рассмотрены понятия двудольного графа, мер центральности (степень вершины, посредничество, близость, собственная центральность), связности и компонент связности. Описаны структурные метрики сети: средняя степень графа, коэффициент кластеризации, средняя длина кратчайшего пути.

Проведен обзор трех ключевых генеративных моделей сетевых структур: модели Эрдеша-Реньи (случайный граф с распределением степеней, стремящимся к распределению Пуассона), модели Уоттса-Строгаца (модель "тесного мира", сочетающая высокий коэффициент кластеризации с логарифмической зависимостью средней длины пути от числа узлов) и модели Барабаши-Альберт (безмасштабная сеть, формируемая через механизмы роста и предпочтительного присоединения, приводящая к степенному распределению степеней).

Описан анализ общественной структуры, в котором сообщество определяется как подмножество узлов, связанных внутри себя плотнее, чем с остальной частью графа. Количественной мерой качества разбиения служит модулярность, предложенная Ньюманом и Гирваном.

Обоснован выбор эмпирической базы данных — репозитория DBLP, представляющего собой открытый, регулярно обновляемый источник библиографических метаданных в области компьютерных наук. Репозиторий функционирует с начала 1990-х годов под руководством Трирского университета (Германия) и аккумулирует структурированную информацию о десятках миллионов публикаций. Важным структурным аспектом данных является

неявная бипартитная модель “автор-публикация”, при переходе к анализу кооперационных связей применяется односторонняя проекция, в результате которой формируется однородный граф соавторства.

Отмечены методологические ограничения использования DBLP: доменная специфика (преимущественно компьютерные науки), потеря контекстуальной информации при проекции, а также проблема идентификации авторов (омонимия, изменение фамилий).

Проведен предварительный прикладной анализ эмпирической базы данных. Рассматриваемый массив содержит один миллион записей. Установлено, что среднее число авторов на одну публикацию составляет 3,25, общее число активных авторов — 957 337. Число авторов с единственной публикацией составляет 544 144, что сразу указывало на высокую степень фрагментации будущей сети. Набор данных охватывает записи за 78 лет (1937-2017), при этом 13 временных диапазонов содержат менее 10 записей. Количество уникальных мест публикации — 4 076, из которых более 1 400 содержат менее 10 записей.

Во втором разделе работы проведен топологический анализ построенной сети соавторства, направленный на выявление узлов централизации и оценку связности графовых компонент.

Исследование точек централизации проведено с двух позиций: через призму мест публикации работ и через анализ авторского состава. Выявлены пять площадок с наибольшей концентрацией статей. На первом месте — “International Conference on Acoustics, Speech, and Signal Processing” с 11 832 публикациями, далее следуют “International Conference on Robotics and Automation” (9 824), “Lecture Notes in Computer Science” (8 206), “International Conference on Image Processing”” (7 611) и “International Conference on Communications” (7 377). Высокая плотность публикаций в определенных источниках свидетельствует о формировании устойчивых научных сообществ вокруг данных площадок.

Анализ коавторских связей показал, что среднее значение количества соавторов на одного исследователя находится в диапазоне от 1 до 20, однако наблюдаются существенные выбросы: у ряда авторов суммарное число уникальных соавторов за весь период наблюдений превышает 1 000 человек. Исходя из теории сложных сетей, такие вершины с высокой долей вероятно-

сти выполняют функцию структурных хабов, обеспечивая связность наиболее крупных компонент графа. Одновременно на общую топологию и связность формируемой двудольной сети критическое влияние оказывают авторы с низкой степенью коавторства (не более пяти связей). Статистика по группе исследователей с минимальным числом соавторов обобщена: 1 соавтор — 17 186 авторов, 2 соавтора — 96 378, 3 соавтора — 151 602, 4 соавтора — 143 819, 5 соавторов — 110 321. На основании приведенных статистических выкладок можно с высокой степенью достоверности утверждать, что исследуемая сеть не является связной в полном объеме: количество компонент связности в ней будет превышать единицу.

Рассмотрена первая пятерка авторов с наибольшим числом соавторов: Wei Wang (1 864), br (1 540), Wei Li (1 403), Yang Liu (1 392), Lei Wang (1 332). Пиковые значения активности коавторства приходятся на период с 2009 по 2017 годы.

Анализ структуры и динамики компонент связности соавторской сети показал выраженную неоднородность структуры: подавляющее большинство компонент содержат не более 50 узлов-авторов. Основную долю малоразмерных компонент составляют изолированные авторы, публиковавшие работы без соавторства. Оценка ежегодного прироста количества компонент связности продемонстрировала динамику, приближенную к экспоненциальной: число вновь появляющихся компонент стабильно увеличивается на протяжении исследуемого периода. Более детальный анализ компонент малого размера (1 и 2 автора) позволил выявить устойчивые паттерны фрагментации сети: количество компонент каждого из указанных размеров также подчиняется экспоненциальной тенденции роста. Подобная однородность динамики для компонент различного размера указывает на системный характер процесса фрагментации, а не на случайные флуктуации.

Кумулятивный анализ динамики компонент связности с учетом временной эволюции показал существенно более сглаженную траекторию роста, аппроксимируемую параболической функцией, в отличие от экспоненциальной динамики ежегодных срезов. Подобное поведение указывает на наличие баланса между процессами фрагментации (появление новых одиночных авторов) и интеграции (встраивание одиночных авторов в существующие компоненты).

В третьем разделе работы проведен эмпирический анализ частотных характеристик лексических единиц на основе контент-анализа названий и аннотаций научных публикаций.

Методология контент-анализа включает три этапа: подготовительный (определение целей и задач, разработка классификатора), исполнительный (кодировка данных в соответствии с инструкцией) и этап обработки собранных материалов (статистическая обработка и интерпретация). Критерием фильтрации авторов являлась их принадлежность к компоненте связности самого объемного размера в сети, что позволило сосредоточить анализ на авторах с определенной публикационной активностью и исключить возможность неправильной интерпретации результатов сходства несвязанных публикаций.

Для технической реализации задействован язык программирования Python и библиотека Natural Language Toolkit (NLTK), в рамках которой применены модули стоп-слов (для фильтрации нерелевантных лексических единиц) и лемматизатор WordNet Lemmatizer (для приведения слов к базовой форме). Результаты статистического подсчета сериализуются в бинарный формат с помощью стандартной библиотеки Pickle.

На первом этапе исследования в корпусе было идентифицировано 351 894 уникальных слов, суммарное количество всех прошедших фильтрацию словоупотреблений составило 53 170 081. Наиболее частотное слово корпуса — “Network” — составляет лишь 0,837% от общего числа словоупотреблений (444 895 вхождений). Несмотря на кажущуюся незначительность доли, в условиях лексического разнообразия свыше 350 тысяч единиц такая концентрация указывает на высокую универсальность данной лексемы.

Для оценки эволюции тематических приоритетов научного сообщества статистика словоупотреблений сгруппирована по годам издания публикаций. Сформированный массив охватывает 77 лет исследовательской активности. В качестве иллюстративного среза рассмотрены 1995, 2005 и 2015 годы. Лидирующие позиции занимают лексемы “Network”, “Time”, “Image”. Динамика частоты употребления данных терминов на всем протяжении рассматриваемого периода демонстрирует высокую степень синхронности: пики и последующие снижения фиксируются в совпадающих временных интервалах, что свидетельствует о скоординированном характере исследовательской активно-

сти. Подобная корреляция позволяет интерпретировать данные лексемы как общие маркеры тематического пространства.

Сравнительный анализ публикационной активности и лексического разнообразия отдельных исследователей показал ожидаемую корреляцию: более активные авторы, как правило, затрагивают более широкий круг задач, требующий разнообразного терминологического аппарата. Зарегистрированное отклонение в показателях Lei Zhang (большой объем текста при относительно меньшем лексическом разнообразии) может указывать на специфический подход к аннотированию, при котором детализация методологии достигается за счет повторения ключевых конструкций. Превышение показателя лексического разнообразия у Wei Li при меньшем количестве публикаций может свидетельствовать о междисциплинарном характере исследований. Траектории как общего объёма текста, так и лексического разнообразия у наиболее продуктивных авторов демонстрируют выраженный экспоненциальный рост после 2000 года, что соответствует начальному этапу их академической карьеры.

В четвертом разделе работы проведен векторный анализ и интерпретация результатов, представляющая собой завершающую стадию исследования. Ключевой научной задачей данного этапа являлось выявление и количественная оценка степени влияния лексического профиля каждого исследователя на семантическую структуру текстов, создаваемых в рамках совместной научной деятельности его соавторами.

Для решения поставленной задачи текстовые данные формализованы через векторную модель представления текстовых корпусов. Каждый текстовый документ — будь то отдельная публикация, совокупность аннотаций автора или полный корпус его работ — представлен в виде точки в многомерном векторном пространстве. Оси (координаты) данного пространства формируются уникальными лексическими единицами предметного словаря, сформированного на предыдущих этапах работы. Размерность пространства равна количеству уникальных словоформ, прошедших процедуру фильтрации и лемматизации. Формально лексический профиль автора A представлен в виде вектора:

$$\vec{v}_A = (w_{A,1}, w_{A,2}, \ldots, w_{A,n}), \tag{1}$$

где n — размерность словаря, а $w_{A,i}$ — количественная мера использования i-го слова в корпусе публикаций автора A.

Проведен теоретический обзор метрик сравнения текстовых векторов: евклидово расстояние, косинусное сходство, коэффициент корреляции Пирсона, дивергенция Кульбака-Лейблера и расстояние Дженсена-Шеннона. На основании представленного обзора для целей настоящего исследования выбрана метрика косинусного сходства в качестве основного инструмента количественной оценки семантической близости лексических профилей авторов. Данный выбор обусловлен следующими методологическими основаниями: (1) инвариантность к масштабу — метрика позволяет корректно сравнивать лексические профили независимо от абсолютного числа публикаций; (2) интерпретируемость — значение косинусного сходства имеет наглядную геометрическую интерпретацию (угол между векторами) и интуитивно понятный диапазон значений $[0, 1]$; (3) вычислительная эффективность — операция скалярного произведения и вычисления норм векторов обладает линейной сложностью $O(n)$ по размерности словаря.

В качестве репрезентативного объекта для детального анализа избран исследователь Wei Wang, занимающий позицию узла с максимальной степенью связности в исследуемой сети соавторства. Количество его кооперационных связей достигает 1 863 уникальных соавторов. Общая размерность пространства составляет 844 980 измерений. Для оптимизации использования оперативной памяти реализован подход разреженного хранения данных: лексический профиль каждого автора кодируется в виде упорядоченной последовательности кортежей формата “(идентификатор лексики, частота встречаемости)”.

Реализован алгоритм попарного вычисления метрики косинусного сходства между лексическим вектором целевого автора и векторами всех 1 863 его соавторов. Результаты систематизированы: пять максимальных значений косинусного сходства составили 0,906; 0,891; 0,881; 0,874; 0,871, а пять минимальных — 0,002; 0,022; 0,038; 0,038; 0,043. Среднее арифметическое значение косинусного сходства по всей выборке соавторов составило 0,3653. Указанный показатель свидетельствует о наличии умеренной степени лексической общности, что согласуется с теоретическими ожиданиями для научных коллабораций: авторы, работающие в смежных проблемных областях, демонстриру-

ют частичное пересечение терминологического аппарата, однако сохраняют индивидуальную стилистическую и предметную специфику.

Для выявления потенциальных структурных взаимосвязей между частотой научного взаимодействия и степенью лексической согласованности проведен корреляционный анализ. Анализ эмпирического распределения показывает, что увеличение числа совместных работ не всегда сопровождается пропорциональным ростом лексического сходства, что указывает на сложную, нелинейную природу формирования научных коллабораций.

Проведено детализированное исследование эволюции лексического состава на примере парных взаимодействий. Рассмотрены три сценария совместной активности: публикация одной работы, двух работ и трех работ. Представлена динамика изменения коэффициента косинусного сходства для пары исследователей, объединенных единственной совместной работой. Анализ траектории показывает, что в период до выхода общей публикации показатель сходства демонстрировал волнообразный характер с локальными колебаниями. В момент издания совместной статьи фиксируется выраженный положительный тренд, свидетельствующий о сближении терминологического аппарата соавторов.

Ситуация, отраженная далее, иллюстрирует эволюцию сходства в парах авторов, имеющих в своем активе две совместные публикации. На начальном этапе картина сохраняет сходство с первым сценарием: фиксируется рост лексической близости. Однако после выхода второй совместной работы траектория меняется, и показатель косинусного сходства начинает демонстрировать поступательное снижение. Данное явление может интерпретироваться как следствие дивергенции научных интересов: по завершении цикла совместных проектов каждый из исследователей возвращается к собственным тематическим приоритетам.

Аналогичная тенденция прослеживается при анализе пар с тремя совместными публикациями. В период активности, охватывающий первые два совместных труда, отмечается стабильный рост коэффициента сходства. Последующее снижение показателя после выхода третьей работы требует осторожной интерпретации: наблюдаемый спад обусловлен не столько изменением исследовательских стратегий авторов, сколько методологическим ограничением выборки — данные за 2016 год и последующие периоды представлены

в базе не в полном объеме.

Подводя итог анализу временной динамики, можно констатировать, что совместная публикационная деятельность оказывает непосредственное влияние на лексическую близость авторов. На начальном этапе сотрудничества наблюдается конвергенция терминологического аппарата, однако по завершении цикла совместных работ тенденция сменяется постепенной дивергенцией.

ЗАКЛЮЧЕНИЕ

В ходе выполнения настоящей работы была достигнута поставленная цель — исследованы закономерности поведения групп в сложных сетях на основе репозитория DBLP и проанализирована взаимосвязь между структурой сети соавторов и лексическим содержанием публикаций.

Для достижения цели были последовательно решены следующие задачи:

1. **Теоретическое обоснование и подготовка данных.** Изучены фундаментальные принципы теории сложных графов, определены ключевые топологические метрики. В качестве эмпирической базы выбран репозиторий DBLP, проведен его предварительный анализ и осуществлена предобработка с удалением нерелевантных записей для обеспечения репрезентативности выборки.
2. **Построение и топологический анализ сети.** На основе данных о соавторстве сформирован двудольный граф. Установлено, что сеть обладает слабой связностью: выявлено значительное число изолированных компонент малого размера (1-2 автора), в то время как основная часть исследователей входит в гигантскую компоненту. Определены вершины с максимальной степенью, выступающие центрами консолидации сетевых кластеров.
3. **Контент-анализ публикаций.** С применением библиотеки NLTK проведена автоматизированная обработка текстовых данных (токенизация, лемматизация, фильтрация стоп-слов). Сформированы лексические векторы для каждого автора, отражающие частотное распределение ключевых терминов в их работах за весь период наблюдений.
4. **Интерпретация результатов.** Вычислено попарное косинусное сходство лексических профилей соавторов. Установлено, что наличие совместных публикаций положительно коррелирует с ростом лексической близости авторов. Динамика сходства демонстрирует скачкообразное увеличение в периоды активного соавторства и постепенное снижение при смене научных интересов или прекращении совместной деятельности.

Основные выводы:

— Сеть соавторов DBLP характеризуется выраженной гетерогенностью распределения связей и наличием множества изолированных микроструктур.

— Лексический состав публикаций служит устойчивым индикатором научной специализации автора и чувствителен к изменениям в тематике исследований.

— Совместная публикационная деятельность способствует конвергенции терминологического аппарата соавторов, что подтверждается статистически значимыми значениями косинусного сходства.

— Увеличение числа совместных работ не всегда сопровождается пропорциональным ростом лексического сходства, что указывает на нелинейную природу формирования научных коллабораций.

— Кумулятивный подход к анализу компонент связности (с учетом преемственности между годами) дает более реалистичную картину эволюции сети по сравнению с погодно-срезовым анализом, демонстрируя баланс между процессами фрагментации и интеграции.

Полученные результаты могут быть использованы для прогнозирования формирования новых научных коллабораций, анализа эволюции исследовательских областей и разработки рекомендательных систем в академической среде.

В качестве направлений для дальнейших исследований предлагается учесть семантическую близость терминов (word embeddings) и расширить выборку за счет привлечения дополнительных библиографических баз данных.

Основные источники информации:

1 Newman, M. E. J. The structure and function of complex networks [Text] / M. E. J. Newman // SIAM Review. — 2003. — Vol. 45, no. 2. — P. 167-256.

2 Complex networks: Structure and dynamics [Text] / S. Boccaletti, V. Latora, Y. Moreno [et al.] // Physics Reports. — 2006. — Vol. 424, no. 4-5. — P. 175-308.

3 Newman, M. E. J. Networks [Text] / M. E. J. Newman. — 2 edition. — Oxford: Oxford University Press, 2018. — ISBN: 9780198805090.

4 Wasserman, Stanley. Social Network Analysis: Methods and Applications [Text] / Stanley Wasserman, Katherine Faust. — Cambridge: Cambridge University Press, 1994. — ISBN: 978-0-521-38707-1.

5 Barabási, Albert-László. Network Science [Text] / Albert-László Barabási. — Cambridge: Cambridge University Press, 2016. — ISBN: 978-1-107-07626-6.

6 Barabási, A.-L. Emergence of scaling in random networks [Text] / A.-L. Barabási, R. Albert // Science. — 1999. — Vol. 286, no. 5439. — P. 509-512.

7 Watts, D. J. Collective dynamics of 'small-world' networks [Text] / D. J. Watts, S. H. Strogatz // Nature. — 1998. — Vol. 393, no. 6684. — P. 440-442.

8 Erdős, P. On random graphs [Text] / P. Erdős, A. Rényi // Publicationes Mathematicae Debrecen. — 1959. — Vol. 6. — P. 290-297.

9 Albert, R. Diameter of the world-wide web [Text] / R. Albert, H. Jeong, A.-L. Barabási // Nature. — 1999. — Vol. 401, no. 6749. — P. 130-131.

10 Newman, M. E. J. Finding and evaluating community structure in networks [Text] / M. E. J. Newman, M. Girvan // Physical Review E. — 2004. — Vol. 69, no. 2. — P. 026113.

11 Fortunato, S. Community detection in graphs [Text] / S. Fortunato // Physics Reports. — 2010. — Vol. 486, no. 3-5. — P. 75-174.

12 Ley, Michael. The DBLP Computer Science Bibliography: Evolution, Research Topics, Future Prospects [Text]. — Springer-Verlag, 2000.

13 Otte, E. Social network analysis: a powerful strategy, also for the information sciences [Text] / E. Otte, R. Rousseau // Journal of Information Science. — 2002. — Vol. 28, no. 6. — P. 441–453.

14 Newman, M. E. J. The structure of scientific collaboration networks [Text] / M. E. J. Newman // Proceedings of the National Academy of Sciences. —

2001. — Vol. 98, no. 2. — P. 404-409.

15 Evolution of the social network of scientific collaborations [Text] / A.-L. Barabási, H. Jeong, Z. Néda [et al.] // Physica A: Statistical Mechanics and its Applications. — 2002. — Vol. 311, no. 3-4. — P. 590-614.

16 Author disambiguation in heterogeneous networks [Text] / J. Han, Y. Sun, X. Yan, P. S. Yu // ACM SIGKDD Explorations Newsletter. — 2016. — Vol. 18, no. 1. — P. 1-14.

17 Bird, Steven. Natural Language Processing with Python [Text] / Steven Bird, Ewan Klein, Edward Loper. — O'Reilly Media, 2009. — ISBN: 978-0596516499.

18 Python Software Foundation. Python Language Reference, Version 3.x [Text]. — 2023. — URL: https://www.python.org/.

19 Krippendorff, K. Content Analysis: An Introduction to Its Methodology [Text]. — 3rd ed. — Thousand Oaks: SAGE Publications, 2013. — ISBN: 978-1-4129-9477-5.

20 Loper, Edward. NLTK: The Natural Language Toolkit [Text] / Edward Loper, Steven Bird // Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. — 2002. — P. 63-70.

21 Python Software Foundation. pickle — Python object serialization [Text]. — 2026. — Online documentation; accessed 2026-04-19. URL: https://docs.python.org/