

# ПРИКЛАДНОЙ СЕТЕВОЙ АНАЛИЗ ДЛЯ РЕШЕНИЯ СОВРЕМЕННЫХ ЗАДАЧ ГОСУДАРСТВА, БИЗНЕСА И ОБЩЕСТВА

МЕТОДОЛОГИЧЕСКИЕ РАЗРАБОТКИ И ПРАКТИЧЕСКОЕ ПРИМЕНЕНИЕ

November 30, 2023

## Содержание

<b>1 Введение</b>	<b>3</b>
1.1 Актуальность исследования . . . . .	3
<b>2 Современные статистические методы для сетевого анализа: возможности и потенциал применения байесовской статистики</b>	<b>4</b>
<b>3 1.2 Акторно-ориентированные стохастические модели для изучения сетевой динамики и социального влияния</b>	<b>13</b>
3.1 Введение . . . . .	13
3.2 Изучение сетевой динамики: перспективы, ограничения и применения SAOM . . . . .	14
3.3 ERGM и tERGM для моделирования динамических сетей . . . . .	16
3.4 Базовая модель SAOM . . . . .	18
3.5 Оптимизация и оценка . . . . .	19
3.6 Вырождение . . . . .	21
3.7 Заключение . . . . .	21
3.8 Новые типы сетей и возможности их анализа: Многосторонние сети . . . . .	22
3.9 Применение современных методов машинного обучения для предсказания связей в социальных сетях . . . . .	24
3.10 Современные подходы в области сетевой кластеризации и блокмоделинга: систематизация и сравнительный анализ . . . . .	33
3.11 Современные подходы в области статистического сетевого анализа и моделирования: модели SIENA, ERGM, tERGM . . . . .	58
3.12 Современные методы анализа неструктурированной текстовой информации: систематизация и сравнительный анализ . . . . .	70
3.13 Внедрение внешних знаний в языковые модели . . . . .	73
<b>4 Возможности развития методики когнитивного интервью</b>	<b>77</b>
4.1 Введение . . . . .	78
4.2 Цели и задачи исследования . . . . .	78
4.3 Практическая значимость исследования . . . . .	79
4.4 Обзор литературы . . . . .	79
4.5 Методы сбора и обработки данных . . . . .	82
4.6 Обсуждение . . . . .	84
4.7 Результаты исследования . . . . .	85
4.8 2.5 Стратегия качественного сетевого анализа: основные подходы, возможности и потенциал применения в прикладных исследованиях . . . . .	87
4.9 3.1.1 Программа «Bib-eLib» для сбора и обработки библиографических данных на русском языке из электронной библиотеки eLibrary . . . . .	103
<b>5 Методологические особенности предобработки данных по российским авторам в Web of Science</b>	<b>108</b>

<b>6 Сравнительный анализ возможностей баз данных Web of Science и eLibrary для анализа библиографических сетей</b>	<b>113</b>
6.1 Введение . . . . .	113
6.2 Обзор литературы . . . . .	115
6.3 Результаты исследования . . . . .	124
6.4 Адаптация методов текстового анализа для извлечения информации графовыми методами. . . . .	139
<b>7 Сравнительный анализ актуальных подходов к анализу неструктурированной текстовой информации (стохастический блокмоделинг, LDA и BERT модели) на примере анализе дискурсов в социальных медиа</b>	<b>143</b>
7.1 Введение . . . . .	143
7.2 Цели исследования . . . . .	144
7.3 Практическая значимость исследования . . . . .	144
7.4 Обзор современных методов тематического моделирования . . . . .	144
7.5 Методология и дизайн исследования . . . . .	146
7.6 Эмпирическая база исследования . . . . .	147
7.7 Результаты . . . . .	148
7.8 Обсуждение и выводы . . . . .	151
7.9 Заключение . . . . .	151
7.10 4.2.1 Библиометрический сетевой анализ коллабораций российских социологов на материалах Web of Science . . . . .	152
7.11 4.2.3 Картрирование научного поля: применение VOSviewer и Biblioshiny на материалах Web of Science . . . . .	163
7.12 5.11.1 Модели управления благотворительными фондами – бенефициарными собственниками бизнес-компаний . . . . .	173
7.13 5.11.2 Изучение гендерной специфики деструкторов руководителей крупных российских компаний . . . . .	178
<b>8 Профессиональные роли журналистов: об исследовательском проекте</b>	<b>187</b>
8.1 Введение . . . . .	187
8.2 Методы сбора и обработки данных . . . . .	188
8.3 5. Обеспечение практического внедрения инструментов в прикладные аналитические и консалтинговые проекты . . . . .	201
8.4 5.3 Разработка и бизнес-применение инструментария для оценивания репутации брендов на основе семантического сетевого анализа . . . . .	203
8.5 Реализация интегрированных подходов . . . . .	204
<b>9 Здоровые и безопасные города: актуальные тренды исследований в научной литературе и социальных медиа</b>	<b>205</b>
9.1 Введение . . . . .	205
9.2 Обзор литературы . . . . .	206
<b>10 Методы сбора и обработки данных</b>	<b>207</b>
<b>11 База данных</b>	<b>209</b>
11.1 Результаты . . . . .	211
11.2 Обсуждение и выводы . . . . .	220
<b>12 5.6 Качественный сетевой анализ в эмпирических исследованиях</b>	<b>221</b>
<b>13 Применение ERGM для анализа конференций</b>	<b>232</b>
13.1 Введение . . . . .	232
13.2 Данные и методы . . . . .	233
13.3 Результаты . . . . .	235
13.4 Заключение . . . . .	239
13.5 5.9 SNA для анализа пользовательского контента: взгляд через призму маркетинга . . . . .	240
13.6 3. Процесс применения SNA для анализа пользовательских отзывов . . . . .	246
13.7 3.2 Анализ основных метрик SNA . . . . .	247

# **1 Введение**

## **1.1 Актуальность исследования**

Сетевой анализ как консистентная исследовательская методология сформировался в 1970-80-е гг., объединив ряд наработок в области социальной психологии, социометрии, социологии, антропологии, экономики, политологии, социальной географии, математики (теории графов) и статистики, а с 2000-х гг. стал разрабатываться также в естественных науках, что привело к появлению науки о сетях (Network science). Сетевой анализ относится к системному уровню анализа и рассматривает эмпирически обозримые отношения в виде сети, состоящей из узлов, связанных направленными или ненаправленными связями различной интенсивности. Предметом исследования выступают глубинные социальные структуры, оказывающие ограничивающее влияние на акторов с разным положением в социальной структуре и неравным доступом к ресурсам.

К настоящему времени в прикладном сетевом анализе разработано большое количество продвинутых методов для анализа различных типов сетевых данных. Использование этих методов позволяет отвечать на множество важных вопросов и решать современные задачи, стоящие перед государством, бизнесом и обществом.

Данный текстовый отчет содержит описание основных результатов работы Международной лаборатории за 2022 год по основным тематическим направлениям деятельности сотрудников. В главе 1 описана история появления и развития сетевого анализа, теоретические положения, определения основных понятий и принципы проведения сетевых исследований, а также представлены основные методологические разработки в сетевом анализе, сделанные в рамках реализации проекта. В главе 2 приведено описание проектов, реализованных сотрудниками лаборатории в 2022 году, где применялись методы, модели и инструменты для сбора, очистки и анализа социальных сетей. Для каждого проекта описаны их цели и задачи, методы сбора и обработки данных, полученные результаты, область их применения и степень внедрения. А вот пример цитирования — [larranaga2013?].

## **2 Современные статистические методы для сетевого анализа: возможности и потенциал применения байесовской статистики**

Обнаружение сообществ в сетях широко изучалось в сетевой науке. Оно привлекло внимание в 1970-х годах, когда Лоррейн и Уайт [245] ввели функциональный мэппинг для получения глобальных сетевых паттернов, таких как структурная эквивалентность. Было предложено множество алгоритмов и методов, направленных на понимание структуры сложных систем. Однако, даже после недавнего оживления этой области, связанного с методом, основанным на спектральной модулярности [285], в исследовательском сообществе уже давно существует понимание того, что различия в простой концептуализации того, что такое сообщество, не являются основанием для существования единого метода для каждого приложения. Хотя не существует единого метода для каждого приложения [125], особенно при различной концептуализации сообществ, важно изучить различные сети и подходы. В данном разделе представлены современные статистические методы на основе байесовской статистики для сетевого анализа, применяемые для обнаружения сообществ в статических и динамических сетях. В разделе описаны особенности изучения развивающихся (динамических) больших разреженных сетей с использованием методов сетевого анализа. Описаны особенности и выявлены преимущества использования байесовского подхода в сетевом анализе для обнаружения сообществ в динамических сетях. Приведены примеры применения байесовского подхода к обнаружению сообществ при разработке больших разреженных сетей в области изучения различных социальных процессов.

Интерес к изучению социальных групп является фундаментальным для социологии. Классики социологии исследовали природу и эволюцию сообществ, сравнивая традиционные и современные формы, концепции Тенниса о *Gesellschaft* и *Gemeinschaft*, Дюркгейма о механической и органической солидарности продолжают оказывать влияние на социологические теории и в настоящее время. Однако также существует набор теорий для более сфокусированного анализа взаимодействий индивидов и, в последствии, анализа социальных сетей. Георг Зиммель [353], Чикагская школа социологии [82], и теоретики социального обмена [178] изучали характеристики социальных групп. Одним из важных понятий является первичная группа - небольшое неформальное сообщество с личным контактом и позитивным отношением [126], и важность этих групп заключается в формировании паттернов социального поведения и ценностных установок по отношению к участникам группы и индивидам за ее пределами [129]. Структура подобных групп была впервые описана Дэвисом [95] в 1941 году на примере женских сообществ в южных штатах США, где он выделил первичных и вторичных участников групп на основе частоты их участия.

К операционализации социальных групп можно подходить по-разному, в зависимости от того, как исследователи рассматривают групповую сплоченность. Существует несколько методов: 1) предполагая полную взаимность связей, когда каждый участник выбирает другого; 2) наличие доступных участников, подключенных через других участников; 3) подчеркивание частых контактов или большого количества соседних узлов в сетевом анализе; 4) изучение внутригрупповой связанности по сравнению с внегрупповой связанностью [413]. Каждый метод опирается на классические концепции и методы анализа социальных сетей, некоторые из которых являются частью современных инструментов обнаружения сообществ. Одним математическим понятием, соответствующим полной взаимности, является клика, определяемая как полный подграф взаимно смежных узлов, при этом никакие другие узлы не имеют такой же

смежности [249]. Однако клики имеют ограничения в реальных социальных сетях, которые часто могут быть разреженными, что затрудняет поиск плотных групп. Кроме того, клики не дают представления о внутренней структуре группы, поскольку все участники имеют одинаковое количество связей и фактически идентичны. Таким образом, клики обладают ограниченным исследовательским потенциалом на практике.

Математические определения сплоченных социальных групп основаны на концепции клики, но с ослабленными свойствами. Одной из таких концепций является  $n$ -клика, члены которой могут относительно легко связываться друг с другом через посредников.  $N$ -клика - это максимальный подграф, в котором наибольшее геодезическое расстояние между любыми двумя узлами не превышает  $n$  [22]. По мере увеличения  $n$  все больше узлов распознаются как часть  $n$ -клики, включая те, которые подключены к остальным максимум через  $n$  посредников. Однако, в отличие от клики,  $n$ -клика менее сплочена. Он может не иметь диаметра размера  $n$ , и даже при  $n=2$  наибольшее геодезическое расстояние внутри  $n$ -клики может быть больше  $n$ . Кроме того,  $n$ -клика может быть не подключен, так как путь между узлами может проходить в обход элементов  $n$ -клика. Для решения этих проблем есть два решения:  $n$ -клан и  $n$ -клуб [274], но они не получили широкого применения в исследованиях [376].

Сплоченная подгруппа, построенная вокруг частых (но не обязательно постоянных) контактов ее членов, отражена в концепциях  $k$ -plex и  $k$ -core. Они были созданы потому, что  $n$ -клики подвержены уязвимости, т.е. исчезновению из-за удаления одного из узлов [351], поэтому было необходимо разработать что-то, что повысило бы общую сплоченность подгрупп. В этом отношении  $k$ -plex определяется как максимальный подграф  $S$  с  $ns$  узлами, которые привязаны к минимуму узлов  $ns-k$ . Иными словами, у каждого узла в  $S$  могут отсутствовать связи максимум с  $k$  членами  $S$ . Степень (количество смежных узлов) для каждого узла в  $S$  равна не менее  $ns-k$  и не может быть больше  $ns-1$ . При  $k = 1$  это дает нам клику, и по мере увеличения  $k$  увеличивается и количество недостающих звеньев. Но поскольку связи охватывают более широкий диапазон, это делает подграф более стабильным. Однако крайне важно установить правильный размер подграфа (количество узлов), чтобы  $k$ -множества не были очень разреженными.  $k$ -ядро — это максимальный подграф, где каждый узел имеет степень не менее  $k$  [349].  $k$ -core является противоположностью  $k$ -plex, потому что в нем указано, сколько связей узел должен иметь с другими элементами своего подграфа, а не сколько у него может не быть. Вот почему  $k$ -ядра также часто являются вложенными, т.е. узлы могут находиться как в  $k$ -ядре, так и в  $(k+1)$ -ядре. Хотя этот подход не приводит к созданию плотных группировок, он служит показателем для общей кластеризации и определения мест потенциальных группировок [349].

Понимание того, как устроено сообщество, включает в себя связи, близость и частоту контактов между членами группы. Множества  $LS$  и  $\lambda$  отражают идею многочисленных и прочных связей внутри сообщества.  $LS$  имеет больше связей внутри своих подграфов, чем за их пределами [350], в то время как  $\lambda$  использует линейную связность для определения минимального количества связей для разъединения узлов. Множества  $\lambda$  являются более общими и отражают уникальное понятие связности. Эти концепции являются основополагающими для изучения сообществ в рамках анализа социальных сетей. Понятие клики было расширено, чтобы охватить близость ( $n$ -клик) и частоту контактов ( $k$ -сплетений и  $k$ -ядер). Комплекты  $LS$  и  $\lambda$  сочетают в себе внутреннюю относительную прочность и малое количество внешних соединений. Эти концепции необходимы для передовых методов обнаружения сообществ в разреженных сетях.

В области выявления сообществ проводится множество исследований по новым и систематизированным методам, и хотя согласованной таксономии не существует, есть некоторая основа классификации. Статичные сети далее изучаются с помощью методов секционирования (partitioning) или традиционной кластеризации, а также методов статистического вывода (statistical inference). Некоторые исследователи рассматривают спектральную кластеризацию отдельно, в то время как другие классифицируют ее как основанную на модульной оптимизации или традиционную кластеризацию. Были предложены различные подходы к классификации, включая нулевые модели, блочные модели, потоковые модели и такие функции, как внутренняя плотность, структурное сходство, динамическое сходство и возможность разделения разделов. Вместо генеалогических классификаций мы фокусируемся на общих исследовательских задачах и признаем примеры, на которых основаны методы. В одном разделе мы обсуждаем статические методы обнаружения сообществ, а в другом - динамические методы, описывая подходы, мотивацию, механику и ситуации для их использования

Сначала определим группу подходов, основной целью которых является разделение сети на подходящее количество сообществ. Подходы, основанные на разрезании (cut-based), направлены на разделение сети на сообщества путем минимизации внутренних связей и максимизации внешних. Это достигается с помощью таких методов, как обрезка дуг (edge cut) и пропорциональная нарезка (ratio cut), обеспечивающая сбалансированное и оптимальное разделение. Спектральная кластеризация - популярный метод, который преобразует сеть в точки в многомерном пространстве с использованием собственных векторов, позволяя идентифицировать сообщества в сложных данных. Однако он не масштабируем для больших сетей или разреженных данных. Также используются методы разделения, такие как k-средние значения и алгоритмы иерархической кластеризации (HCA). HCA работает сверху вниз или снизу-вверх для обнаружения иерархических сообществ, но полагается на правильное расположение узлов. Алгоритмы Гирвана-Ньюмана (GN) устраниют границы с высокой степенью связанности для поиска сообществ, но могут быть дорогостоящими с точки зрения вычислений. Эти подходы практичны для различных данных, особенно для исследовательских целей, но могут плохо работать с большими или разреженными наборами данных. Кроме того, они не обладают высокой масштабируемостью.

Поиск среза минимальной дуги сам по себе может не дать удовлетворительных результатов, поскольку он игнорирует внутренние связи подграфов. Чтобы решить эту проблему, проводимость (conductance) была введена как мера среза дуг (edge cut) относительно объема подграфа [192]. Проводимость обеспечивает локальное развертывание и снижает вычислительную сложность [331]. Другой важной концепцией является модулярность, которая количественно определяет количество дуг внутри групп по сравнению со случайной сетью [287]. Максимизация модулярности позволяет идентифицировать сообщества в сети. Спектральная и жадная (greedy) оптимизация — это обычно используемые методы оптимизации модулярности. Спектральная оптимизация использует собственные векторы и собственные значения матрицы модулярности. Жадная оптимизация, такая как алгоритм Louvain, постепенно увеличивает модулярность за счет объединения кластеров. Однако эти алгоритмы могут создавать несвязанные сообщества [188] и в попытке создать не пересекающиеся кластеры, произвести неинтерпретируемые результаты в случае, если в сетях пересекающиеся сообщества, что свойственно реальным сетям [397]. Алгоритм Leiden решает эти проблемы с помощью процедуры быстрого локального перемещения [397]. Несмотря на популярность, методы оптимизации модулярности имеют ограничения,

и исследователи ставят под сомнение их применимость к реальным сетям [124]. В целом, методы, основанные на кластеризации, такие как Louvain и Leiden, широко используются из-за их эффективности в крупномасштабных сетях, но они могут давать неубедительные результаты.

Существуют подходы к выявлению сообществ, основанные на статистическом выводе и проверке гипотез, которые позволяют исследователям тестировать различные типы сообществ в сети. Традиционные методы предполагают плотные и эквивалентные группы, но при анализе социальных сетей бывают случаи, когда узлы могут не группироваться на основе сходства. Могут существовать различные типы сетевых структур, такие как диссортативные и структуры ядро-периферия. Стохастический блокмоделинг (SBM) - это метод, используемый для тестирования конкретных структур сообщества в сети [413], [172]. Он перестраивает связи на основе идеальных блоков, представляющих различные типы сообществ. Цель состоит в том, чтобы свести к минимуму разницу между эмпирической структурой сети и идеальной моделью при оценке соответствия модели. SBM широко используется в SNA и в качестве эталона для методов выявления сообществ [125], [188]. Однако у классического SBM был недостаток, поскольку он не учитывал степень неоднородности в реальных сетях. Для решения этой проблемы был разработан SBM с поправкой на степень (DCSBM) [193], но для такого алгоритма требуется заранее указать количество сообществ, чтобы избежать переобучения модели [125]. Методы выбора модели, такие как иерархическая вложенность SBMS, были предложены для улучшения сжатия данных [302]. Методы, основанные на SBM, являются мощными для выявления сообществ, выступая в качестве генеративных моделей и ориентиров [91]. Однако они сопряжены с более высокими вычислительными затратами и более длительным временем выполнения. Подробности SBM и его байесовского варианта будут рассмотрены позже.

Сообщества обычно рассматриваются как статичные структуры, но также важно учитывать и их динамику. Динамические модели помогают понять функции и временные сообщества внутри сложных систем [331]. Модели, основанные на случайном блуждании [310], идентифицируют сообщества, в которых «случайное блуждание» (random walker) попадает в ловушку, обходя все узлы в этом регионе, и затем эти сообщества группируются с использованием традиционных методов или методов, основанных на разрезе (cut-based) [425]. Однако в первоначальной формулировке этот метод очень требователен к вычислениям и не может быть использован в больших сетях [125]. Подходы, основанные на теории информации, заменяют случайного ходока кодовыми словами для описания структур сообщества и минимизации длины описания. Алгоритм Infomap [332] вычисляет минимальную длину описания бесконечного случайного блуждания и обладает способностью обнаруживать иерархические [333] и перекрывающиеся сообщества [116] и даже воздействовать на сохраненную память о предыдущих состояниях [306]. Существует еще один метод - spin glass, название которого отсылает к моделированию поведения неупорядоченных материалов, таких как стекло, полимеры и сверхпроводники, где большое количество частиц, называемых “спинами”, случайным образом взаимодействуют. В модели spin glass [282] структура сообщества представлена в виде конфигурации спинов, и цель состоит в том, чтобы минимизировать энергию путем согласования спинов со структурой сообщества. Тем не менее алгоритмы случайного блуждания показали лучшую производительность, чем модели spin glass [320]. Динамическое обнаружение сообществ относительно игнорируется [331], с акцентом на статические подходы и фиксацию динамики с помощью моментальных снимков структуры сети. Однако для обнаружения перекрывающихся сообществ и поддержания согласованности с течением времени был

разработан алгоритм динамического байесовского детектора перекрывающихся сообществ (DBOCD).

Для исследования социальных сетей актуальным является вопрос изучения больших разреженных динамических сетей, что предполагает наличие двух качеств сети – «разреженных» и «динамических». Сеть называется разреженной, если в ней гораздо меньше связей, чем возможный максимум [33]. Динамическая сеть рассматривается как функция времени [247], что подразумевает, что для каждого момента в заданном периоде существует граф и изучается множество графов. Следовательно, большая разреженная динамическая сеть может быть определена как подмножество сетей временного масштаба с большим числом узлов и гораздо меньшим количеством связей, чем максимально возможное число.

Причина изучения таких объектов заключается в том, что разреженные сети на самом деле чрезвычайно распространены в реальной жизни, в отличие от плотных модели [33], например, в нейробиологии нейронная система содержит огромное количество нейронов, но связи между ними немногочисленны, что затрудняет анализ. Динамическая характеристика сети также важна, поскольку, изучая эволюцию взаимоотношений узлов и изменения в структуре, исследователи смогли бы отслеживать распространение информации и выявлять особенности построения сообщества.

Ключевой работой, которая обеспечивает основу для математических расчетов в области изучения динамических разреженных сетей, является “Моделирование динамических сетей с разреженными и слабыми связями во временном масштабе” [75]. В книге рассматриваются такие темы, как методология моделирования в масштабе времени, разделение во временном масштабе, преобразование траектории и многое другое. Книга полезна для анализа крупномасштабных систем, таких как энергосистемы, и выявления важных узлов в сети.

Существует несколько методов анализа больших разреженных динамических сетей. В первую очередь стоит рассмотреть алгоритмы анализа графов. Эти алгоритмы используются для анализа структуры сети и выявления закономерностей и взаимосвязей между узлами. Примеры алгоритмов анализа графов включают поиск в ширину, поиск в глубину, кратчайший путь, обнаружение цикла, минимальное связующее дерево и раскраску графа [141].

Теория динамических графов. Этот метод моделирует и анализирует сети как двухкратные разреженные динамические графовые сети с кластерами, представляющими связную подсистему [266]. Применение теории графов позволяет формализовать определения макросостояния системы, микросостояния на микроуровне и динамических структурных изменений [144]. Для классической и динамической теорий графов на первый план выходят различные проблемы. Более конкретно, в классической теории графов основной задачей оптимизации является поиск подграфа или связующего дерева с определенными характеристиками, такими как минимальный вес. В теории динамических графов ключевой проблемой является установление взаимосвязи между оптимационными решениями на разных графах. Это позволяет нам определить наследование в классе динамических графов с общими правилами перехода. Затем мы можем связать наследственное свойство с операциями перехода в траектории динамического графа. Если мы установим эту связь, мы сможем достичь программируемой самоорганизации и получить гарантированные унаследованные структурные свойства и характеристики динамических графов [210]. Спектральные алгоритмы [354], основанные на матричных представлениях сетей, часто используются для обнаружения сообществ, но классические спектральные методы, основанные на матрице смежности и ее вариантах, терпят неудачу в разреженных сетях. Недавно были внедрены новые спектральные методы, основанные на случайных блужданиях (random walks) без обратного отслеживания, которые

успешно обнаруживают сообщества во многих разреженных сетях. Разбиение спектрального графа на самом деле представляет собой семейство методов. Эти методы зависят от собственных векторов матрицы Лапласа или ее родственников в графе. В зависимости от способа разбиения графика спектральные методы можно разделить на два класса. Первый класс использует ведущий собственный вектор графа Лапласиана для двойного разбиения графа. Второй класс подходов вычисляет k-образное разбиение графа с использованием нескольких собственных векторов [334]. Однако стандартные методы, основанные на матрице смежности и связанных с ней матрицах, не работают для очень разреженных сетей, которые включают в себя множество сетей, представляющих практический интерес. В качестве решения этой проблемы недавно было предложено вместо этого сосредоточиться на спектре матрицы без обратного отслеживания - альтернативном матричном представлении сети, которая демонстрирует лучшее поведение в разреженном пределе с несколько иным определением, чтобы обладать желаемыми свойствами, особенно в общем случае сетей с широкой степенью распределения [286].

Вычисления с разреженной матрицей для динамической централизации сети могут быть использованы для получения нового алгоритма вычисления зависящей от времени централизации, который работает с разреженной версией матрицы динамической коммуникабельности. Таким образом, требования к вычислениям и хранилищу сводятся к требованиям разреженной статической сети в каждый момент времени [30]. «Поскольку зависящая от времени пограничная структура обычно позволяет информации широко распространяться по сети, естественная сводка разреженных, но динамичных парных взаимодействий, как правило, принимает форму большой плотной матрицы. По этой причине вычислительные узловые центры для зависящей от времени сети могут быть чрезвычайно дорогостоящими как с точки зрения вычислений, так и с точки зрения хранения; гораздо более дорогостоящими, чем для отдельной статической сети» [30].

Методы пертурбации используются для выделения функциональных аспектов созданной сети, таких как динамика в данной сети. В случае, если исследователи предполагают наличие случайного шума в данных, при выполнении статистического анализа, они многократно пертурбируют и группируют данные, а затем агрегируют результаты [267].

Подходы группы Лассо (Lasso-type approach), зачастую используются в методах сетевого анализа, таких как байесовские подходы и спектральные алгоритмы. Байесовская регрессия Лассо — это тип регрессионного анализа, который выполняет как выбор переменных, так и регуляризацию с целью повышения точности прогнозирования и интерпретируемости результирующей статистической модели [395].

Подходы группы Лассо также используются при оценке и выборе переменных в рамках модели единого индекса [72]. Спектральные алгоритмы, основанные на матричных представлениях сетей, часто используются для обнаружения сообществ, но классические спектральные методы, основанные на матрице смежности и ее вариантах, терпят неудачу в разреженных сетях. Метод «Сетевого Лассо» недавно был адаптирован для наборов данных с сетевой структурой, и было показано, что он может быть точным при определенных условиях, зависящих от базовой структуры сети и набора выборок.

Обнаружение сообществ в больших разреженных сетях является сложной задачей, и традиционные методы зачастую не справляются с такими данными. Байесовский подход приобрел известность благодаря своей способности решать эти проблемы, предоставляя вероятностную структуру, моделирующую неопределенность и включающую предварительную информацию. Он использует

теорему Байеса для обновления убеждений, основанных на новых доказательствах. Байесовские методы широко используются в статистике, сетевом анализе, машинном обучении и искусственном интеллекте. По своей сути байесовский подход основан на теореме Байеса, которая представляет собой математическую формулу, описывающую, как обновлять наши убеждения или вероятности относительно события, когда становятся доступны новые доказательства или информация.

Байесовская вероятность - это математическая основа для обновляющихся убеждений перед лицом новых доказательств. Он объединяет предварительные знания, представленные априорными вероятностями, с наблюдаемыми данными, измеряемыми с помощью вероятностей, для вычисления апостериорных вероятностей. Предварительные данные выражают первоначальные убеждения относительно гипотезы, в то время как вероятности количественно определяют поддержку, предоставляемую данными. Обновляя априорные значения с использованием теоремы Байеса, мы можем получить апостериорные вероятности, которые включают в себя как предшествующие знания, так и новую информацию, что позволяет более точно представлять убеждения [221]. Байесовский подход отлично подходит для обработки неопределенной или неполной информации. Он позволяет систематически обновлять убеждения по мере появления новых свидетельств. Он широко используется в статистике, машинном обучении и сетях. При обучении с учетом разреженности мы фокусируемся на байесовском подходе [385]. Это помогает сделать вывод о структуре сети и динамике с течением времени, учитывая разреженность. Обнаружение сообщества включает в себя поиск групп узлов со схожими схемами подключения. Байесовские методы справляются со сложностями динамических и разреженных сетей. Они дают ценную информацию, учитывая неопределенность и разреженность данных. Байесовский вывод рассматривает вероятности как меру неопределенности и обновляет их новыми данными. Это позволяет принимать рациональные решения и делать прогнозы, сочетая предварительные знания с наблюдаемыми фактическими данными.

Во многих исследованиях изучался байесовский вывод для выявления сообществ. Они используют вероятностные модели для обработки заданий сообщества, структуры сети и временной эволюции. Байесовские априорные значения кодируют убеждения о размерах сообщества, вероятностях связей и динамике. Статья «Байесовское обнаружение сообществ» [299] посвящена применению байесовского вывода к стохастическому блокмоделингу (SBM) для сетевого анализа, в частности для обнаружения сообществ. Байесовские методы восстанавливают метки классов в SBM, предлагая вероятностный подход. В статье приведены теоретические результаты, подтверждающие высокую согласованность байесовского апостериорного метода при обнаружении сообществ, подчеркивая его надежность. Это предполагает проведение будущих исследований по включению предварительных знаний и оценке числа сообществ с использованием байесовских методов. Байесовский вывод улучшает обнаружение сообщества в SBM [299].

В статье «Последовательное байесовское обнаружение сообществ» представлен байесовский подход к обнаружению сообществ в развивающихся больших разреженных сетях, включающий ковариаты. Это решает проблему использования сетевой структуры и ковариационной информации для обнаружения. Предлагаемый байесовский SBM включает ковариаты посредством предварительного случайного разбиения, зависящего от ковариат, выражая их влияние на членство в кластере. Примечательно, что он узнает количество сообществ на основе данных без предварительного знания [168].

Байесовские методы превосходят остальные в выявлении структурных свойств динамических

сетей и моделировании развивающихся взаимосвязей. В статье “Обнаружение сообществ в сетях без наблюдаемых связей” [168] представлена байесовская иерархическая модель для обнаружения сообществ в данных временных рядов без наблюдаемых связей. Этот целостный подход включает в себя байесовский вывод для сравнения моделей и выбора оптимального масштаба сообщества. Он решает вычислительные задачи, проблемы неопределенности и многомасштабного обнаружения. Иерархические байесовские модели интегрируют различные уровни структуры, временные зависимости и свойства сети в рамках единой структуры [145].

В статье «Байесовский подход к идентификации разреженной динамической сети» описывается использование динамических байесовских сетей (DBNs) для идентификации разреженных динамических сетей. DBNs — это вероятностные графические модели, которые фиксируют взаимосвязи между переменными с течением времени [72]. DBNs расширяют байесовские сети для работы с динамическими данными и вероятностными зависимостями. Они находят применение в различных областях, включая обработку языковых данных и финансовое моделирование, для понимания временных взаимосвязей. Более того, байесовские методы могут включать в себя априорные данные, вызывающие разреженность, для идентификации как динамических, так и отсутствующих связей в разреженных сетях. Это предпочтение более простых структур помогает обнаружить значимые взаимосвязи и избежать переобучения [72].

Более того, на основе байесовских методов было предложено множество новых подходов и решений. В статье «Байесовский подход к идентификации разреженной динамической сети» авторы представляют два байесовских подхода, Stable-Spline GLAR (SSGLAR) и Stable-Spline Exponential Hyperprior (SSEH), оба из которых способствуют разреженности при выборе модели и оценке импульсного отклика. Принимая байесовскую точку зрения, можно получить ту же формулировку, моделируя компоненты  $\varphi$  как независимые гауссовские случайные величины. Авторы предлагают байесовскую модель для разреженной идентификации, где распределения вероятностей представляют неопределенность.

Байесовские методы предлагают способ оценки неопределенности с помощью апостериорных распределений. Это особенно важно при выявлении отсутствия динамических связей, поскольку позволяет исследователям не только заявить об отсутствии доказательств наличия связи, но и количественно оценить свою уверенность в этом утверждении. Эта информация ценна для принятия решений и проверки гипотез.

В статье «Обнаружение сообществ и их эволюция в динамических социальных сетях — байесовский подход» [427] предлагается динамическая стохастическая блочная модель для анализа сообществ и их эволюции в единой вероятностной структуре. В методологию интегрирован байесовский подход, использующий байесовский вывод для оценки апостериорных распределений параметров. Это повышает устойчивость модели к помехам в данных и фиксирует неопределенность в значениях параметров. Использование байесовской обработки является ключевой особенностью, которая обеспечивает вероятностную основу для выявления сообществ и анализа эволюции [427].

Динамические сети также имеют переменные прибытия и выбытия узлов. Байесовское моделирование обрабатывает это с помощью вероятностных моделей для этих событий. Предыдущие распределения используются для определения вероятности присоединения новых узлов к сети или выхода из нее на каждом временном шаге. Байесовский вывод обновляет назначения сообществ и параметры модели по мере изменения узлов [427].

Байесовская структура обычно используется для оценки гиперпараметров. Распределения правдоподобия, основанные на модели, включают гауссовские распределения для измерений с шумом [405].

Гауссовые процессы - распространенный инструмент байесовского моделирования - дают оценки и количественную оценку неопределенностей в сценариях с шумом. Они используются для задач регрессии и оценки. В статье «Байесовская оценка импульсных откликов» наиболее вероятный импульсный отклик определяется по наблюдаемым данным. Байесовские подходы включают регуляризацию, чтобы сбалансировать сложность и подгонку данных. В гауссовых процессах выбор ковариационных функций и параметров регуляризации управляет плавностью и регулярностью оцениваемых функций (импульсных откликов). Байесовские методы включают неопределенность и регуляризацию в идентификацию системы, что соответствует подходу, описанному в статье [72].

На практике исследователи часто сталкиваются с тем, что узлы сети группируются в определенные тесные сообщества с преобладанием внутренних связей внутри кластеров над связями между кластерами. Исследователи отмечают, что в сетевом анализе важно эксплицитно моделировать структуру сообществ рассматриваемых сетей. Так, исследуя 16 эмпирических сетей, начиная от классических «Клуба каратэ Захарии» и дельфинов-афалин из «Doubtful Sound» до футбольных и энергосетей, Мёруп и Шмидт показывают, что структура сообществ возникает из всех данных эмпирическим путем и предлагают алгоритм для ее моделирования, опирающийся на байесовскую статистику [279].

Байесовские методы к обнаружению сообществ также могут опираться на классическое понятие ассортативности в формировании связей между узлами сети. В частности, исходя из этого определения, был разработан новый алгоритм стохастического блокмоделинга, который использует логистическую регрессию с поправкой на характеристики узлов. С помощью данного алгоритма авторы выявили сообщества в публикациях политической блогосферы США, а также обнаружили сообщества исходя из политической литературы, которую одни и те же пользователи чаще всего приобретают на Amazon [179].

Развитием байесовского подхода к обнаружению сообществ также является алгоритм неотрицательного матричного разложения (non-negative matrix factorization, NMF). Данный алгоритм позволяет получить вероятностные оценки принадлежности узлов к тому или иному сообществу, а также протестировать конкретное, смоделированное разложение на определенных бенчмарках. Алгоритм применялся, в числе прочего, для анализа сетей взаимодействия между нематодами *Caenorhabditis elegans*; джазовыми музыкантами; студентами одного из университетов в Facebook [315].

Подводя итог, мы можем выделить следующие преимущества Байесовского подхода к обнаружению сообществ:

1. байесовские методы эффективно обрабатывают разреженные данные, обеспечивая значимые результаты при ограниченных наблюдаемых взаимодействиях;
2. байесовские методы автоматически уравновешивают соответствие модели и ее сложность, предотвращая переобучение;
3. байесовский подход предлагает более широкие интерпретации и количественную оценку неопределенности с помощью целых апостериорных распределений;
4. байесовские методы превосходно отражают эволюцию сообщества в динамических сетях;
5. благодаря включению нескольких типов данных в байесовскую структуру повышается точность

- обнаружения сообщества, что дает всестороннее представление о поведении сети;
6. байесовский подход обеспечивает непротиворечивые результаты, согласованные с лежащими в их основе допущениями.

В заключение, применение байесовских методов в сетевом анализе позволяет лучше понять структуру сообщества и его эволюцию с течением времени, решая проблемы неопределенности, разреженности и временной динамики. Они являются ценным инструментом в сетевом анализе.

### **3 1.2 Акторно-ориентированные стохастические модели для изучения сетевой динамики и социального влияния**

#### **3.1 Введение**

В контексте развития сетевого анализа исследования сетей в динамике становятся все более значимыми для понимания сложных взаимосвязей. Осознавая эту потребность, исследователи прибегают к разработке новых методологий для анализа и построения сетей. В этом контексте нельзя не вспомнить про акторно-ориентированные стохастические модели (Stochastic Actor-Oriented Models, SAOMs), представляющие собой одно из наиболее развивающихся и перспективных средств анализа механизмов социального развития, взаимосвязей и эволюции различных сетей. В связи с этим ученые из различных областей, таких как социология, экономика, эпидемиология и коммуникационные исследования, первоочередно прибегают к использованию данного аналитического инструмента для понимания сложных взаимосвязей между акторами.

По сути, SAOM выступает в роли призмы, через которую исследователи могут расшифровать сетевую динамику, раскрывая глубинные процессы, определяющие эволюцию сети. Преодолевая разрыв между наблюдаемым и ненаблюдаемым, модель предоставляет ценный инструмент для сетевых аналитиков, стремящихся выявить скрытые закономерности и механизмы, управляющие социальными взаимодействиями и сетевыми структурами.

В данной работе мы рассматриваем основные принципы и области применения этой методологии. Мы стремимся выяснить отличительные особенности SAOM и их значимость в области стохастического анализа сетей. Проведя сравнительный анализ SAOM и временных экспоненциальных моделей случайных графов (TERGM), мы подчеркнули сильные стороны и уникальный вклад SAOM в раскрытие динамики сетей. Наконец, мы также рассматриваем процесс работы алгоритмов, которые позволяют оценить качество подобных моделей, включая такие статистические показатели как оценка адекватности модели (Goodness of Fit). Кроме того, мы приводим релевантные примеры эмпирических работ, которые позволяют напрямую увидеть практическую значимость методологии в современных исследованиях и дальнейший потенциал для разработки сложных сетевых феноменов.

Цель: Провести сравнительный анализ методологий применения Акторно-ориентированных стохастических моделей (SAOMs) для изучения сетевых динамик и социального влияния.

Задачи: 1. Дать характеристики применения акторно-ориентированных стохастических моделей (Stochastic Actor Oriented Models, SAOM) как одного из направлений развития подходов к анализу динамических сетей. 2. Сравнить акторно-ориентированные стохастические модели (SAOMs) и

темпоральные экспоненциальные модели случайных графов (TERGMs) для изучения динамических сетей.

3. Описать алгоритмы для оценки качества и адекватности акторно-ориентированных стохастических моделей.

### **3.2 Изучение сетевой динамики: перспективы, ограничения и применения SAOM**

Хотя SAOM все еще является развивающимся методом сетевого анализа, его уже успешно применили в различных областях: от анализа небольших сетей дружбы подростков до политологического анализа транснациональных союзов. Мы сделаем обзор некоторых из наиболее цитируемых работ в нескольких научных областях, а также предложим способы применения этого подхода в работе ANR-Lab.

#### **3.2.1 Применения**

**Дружба и влияние сверстников** Применения SAOM породило довольно много исследований сетей с относительно небольшим количеством вершин, отражающих прежде всего дружественные взаимодействия. В них исследовалась reciprocity и гомофилия как каналы социального влияния среди сверстников на такие явления, как курение, употребление веществ, ожирение и т. д. Среди подобных исследований: эффекты пола на распределение индивидуальных характеристик в сетях дружбы [399]; актуальные проблемы курения и употребления алкоголя через призму гомофилии: влияния сверстников [97] [208] [182] [208] [208] [339] и родительского примера [265].

Кроме того, на пересечении медицинской социологии и сетевого анализа, существуют также исследования, применяющие SAOM, для выявления социального влияния на вероятность заболеваний, например, СДВГ [28], подростковой депрессии [433], образа жизни и ожирения [96]. В этой области также существуют работы о социальном влиянии этнического самоопределения [191], лидерских динамик [264], религии [214] и владения оружием [103] на дружбу и процесс выбора друзей.

**Библиометрический анализ** Другим перспективным направлением сетевого анализа социальных сетей, в котором применяется модель Siena, является библиометрический анализ. Наши коллеги по ANR-Lab А. Ферлигой и Л. Кроннегер, в соавторстве с создателем моделей Т. Снайдером, провели впечатляющий анализ научного сообщества и динамик соавторства в Словении с 1996 по 2010 год [122], а также провели дополнительную работу на лучших данных, указав на важность институциональных контекстов на среду работы ученых, его не-механическую природу [219].

Иновации в научных исследованиях, проанализированные с помощью SAOM, уделяют внимание гендерной гомофилии [252], а также сетевым динамикам более молодых областей науки [402], роли административных ресурсов университета на паттерны коллaborации [327], коллаборациям между университетом и индустрией [69] и диффузии инноваций [237].

**Политические науки** Одной из наиболее заметных областей исследований, в которых применяется SAOM, является политология и исследования законодательства. Политические акторы и связи между ними оказались исключительно подходящими для этой модели и позволили провести широкий спектр исследований. Во-первых, это исследования по определению и изменению паттернов коллaborации среди законодателей [185], использующие влияния рисков/ресурсов на принятие решений [45].

В поле исследований также входят статьи по анализу дипломатических связей между странами [206]; международной кооперацией и работой международных союзов в связи с проблемами координации [205] и предсказания будущего Европы как структуры транснациональной сети [393].

Лонгитюдный подход к сетям также полезен при анализе распространения правил и законов через институты, регионы и страны (например, законы вокруг международной торговли [273], а также коэволюции доступа к диджитал-инструментам, демократии и торговых связей [322]. Наконец, SAOM может быть применен для анализа политического действия [319] и исторического моделинга событий [55].

### 3.2.2 Ограничения

SAOM требует больших теоретических оснований, чем TERGM, и модель сочетает как социологические, так и статистические методы [229]. В то же время прочная теоретическая база модели может выступать и ограничением. Для начала, идея о том, что любые изменения в сети происходят исключительно *последовательно*, а не одновременно не позволяет использовать данные из e-mail сообщений, электрических систем, комплексных сетей и других ситуаций, когда “социологическая рамка не соответствует реальности” или она не может быть полностью верифицирована [230]. Из-за этого, наложение новых теоретических предположений на сеть невозможно без предварительной проверки базовых предположений SAOM.

Кроме того, этот метод предполагает, что каждый актор размышляет о своих действиях в одинаковой логике, что удобно для статистического упрощения расчетов, но не всегда соответствует модели с акторами, которые обладают разными классами и мотивами [62]. При этом, нельзя отрицать, что хотя процесс и ускоряется, работа с моделью требует больших временных затрат [362], а получение необходимых данных высокого качества более затратно, в том числе и финансово [375].

### 3.2.3 Перспективы

Перед исследователями стоят еще много методологических проблем, в том числе работа модели с коррелирующими рандомизированными эффектами [62], а также разработка моделей с необнаруженной гетеронормативностью между акторами, которые позволят применить подход к более крупным сетям [363].

Наконец, мы считаем, что SAOMs могут быть применены к нескольким направлениям исследований российского общества. До сих пор только в образовательной сфере были созданы исследования, направленные на выявления факторов академического успеха и поведения учеников с помощью SAOM [455], [456]. Мы предлагаем несколько потенциальных направлений развития.

Библиометрический анализ российского научного сообщества, которым занимается одна из исследовательских групп ANR-Lab уже использует лонгитюдные данные [200], [262]. Более того, в нескольких проектах мы также обладаем крупными датасетами библиометрических данных из Web of Science, которые могут быть в разрезе по отдельными направлениям или институциям проанализированы с помощью SAOM. Кроме того, кажется перспективным анализ открытых судебных данных, а также коллaborаций в законодательных инициативах и торговых международных договорах. В данном случае, наиболее сложным этапом работы был бы сбор и предобработка данных.

### 3.3 ERGM и TERGM для моделирования динамических сетей

Темпорально-стохастический подход открывает большие перспективы в различных областях, включая социальные науки, эпидемиологию, коммуникационные сети и т.д. Наукометрия, включающая количественный анализ научной литературы, коллaborаций и распространения знаний, все больше признает ценность анализа временных сетей [21, 219, 404]. В наукометрии временной стохастический подход представляет собой мощную призму, через которую исследователи могут изучать эволюционирующую ландшафт научных коммуникаций, распространения знаний и сетей сотрудничества. Временная перспектива крайне важна для понимания того, как изменяются сетевые структуры, формируются взаимоотношения и протекают информационные потоки в динамичном мире. Во временном стохастическом подходе предполагается, что сетевые данные могут наблюдаться и измеряться в различные моменты времени. Наблюдения являются не изолированными, а взаимосвязанными – они образуют последовательности, содержащие ценную информацию об эволюции сети.

При работе с кросс-секционными/панельными данными исследователи часто сталкиваются с вопросом выбора между SAOM и TERGM. Согласно предыдущим исследованиям, эти две модели часто дают существенно различающиеся оценки параметров, несмотря на использование практически неразличимых конфигураций моделей [233]. Прежде чем перейти к рассмотрению различий, целесообразно выяснить общие черты, присущие этим моделям.

Прежде всего, следует отметить поразительное сходство математических основ SAOM и TERGM, которые возникли на базе ERGM. Математическое определение вероятности наблюдения  $N$  в базовой модели ERGM выглядит следующим образом:

$$P(N, \theta) = \frac{\exp\{\theta' h(N)\}}{\sum_{N^* \in \mathcal{N}} \exp\{\theta' h(N^*)\}},$$

где  $\theta$  – вектор вещественновзначных параметров (различные значения которых дают те или иные распределения из семейства);  $h(N)$  – вектор статистик наблюданной сети (например, число связей или число треугольников);  $N^*$  – один из элементов  $\mathcal{N}$ .

Для простоты интерпретации разобьем уравнение на четыре части: 1.  $h(N)$  отражает статистики сети; 2.  $\theta$  содержит эффекты; 3.  $\exp\{\theta' h(N)\}$  придает положительный вес наблюданной сети  $N$ . 4.  $\sum_{N^* \in \mathcal{N}} \exp\{\theta' h(N^*)\}$  нормализует все возможные конфигурации  $N$  в  $\mathcal{N}$ .

Как определено в работе Лейфельда и Кранмера, TERGM развивают идею, заложенную в ERGM [232]. Они определяют вероятность сети на текущем временном шаге  $t$  как функцию не только суммы подсчетов подграфов текущей сети, но и предыдущих сетей до временного шага  $t - K$ :

$$P(N^t | N^{t-K}, \dots, N^{t-1}, \theta) = \frac{\exp(\theta^T h(N^T, N^{t-1}, \dots, N^{t-K}))}{c(\theta, N^{t-K}, \dots, N^{t-1})}.$$

При этом предполагается, что статистические показатели, полученные на основе связей между временем  $t - K$  и временем  $t$ , эффективно отражают присущие сети зависимости в момент времени  $t$ . Эта простая идея лежит в основе TERGM. В знаменатель этой формулы входит нормирующая константа, аналогичная той, что используется в ERGM. На следующем этапе определяется вероятность, связанная с временным рядом сетей, путем вычисления произведения всех временных периодов:

$$P(N^{K+1}, \dots, N^T | N^1, \dots, N^k, \theta) = \prod_{t=K+1}^T P(N^t | N^{t-K}, \dots, N^{T-1}, \theta).$$

Это представляет собой простое расширение ERGM на последовательность сетей. Для учета временных зависимостей между последовательными временными шагами вводится статистика сети  $h$ , позволяющая включать в анализ временной аспект. Лейфельд и коллеги предлагают исчерпывающее рассмотрение этого вопроса [232].

В основе SAOM лежит стохастический процесс с непрерывным временем, служащий генеративной моделью для отображения временной эволюции сетей. В отличие от SAOM, TERGM функционирует в основном как общая модель, охватывающая состояния сети на нескольких временных отрезках. Как правило, TERGM параметризуется таким образом, чтобы включить временную динамику в качестве объясняющего фактора в модель. SAOM же фокусируется на моделировании *преобразований*, происходящих между этими временными точками, а не на прямом моделировании конечных результатов, как это происходит в TERGM. Такой микроуровневый подход делают SAOM теоретически более обоснованным для построения моделей, отражающих динамику сети. Оцениваемые параметры в SAOM имеют важное значение, поскольку они представляют собой вклад в так называемую целевую функцию. Как будет показано далее в отчете, оценка параметров SAOM осуществляется с помощью (обобщенного) метода моментов, использующего алгоритм стохастической аппроксимации Роббинса-Монро [364].

В дополнении к этому, SAOM и TERGM являются развивающимися подходами к моделированию, каждое из которых потенциально может предложить определенные преимущества в конкретных контекстных сценариях. Например, SAOM демонстрирует возможность оценки множества взаимосвязанных сетевых процессов в рамках единой модели [348]. Несмотря на возможность применения ERGM к многоуровневым, мультиплексным и мультиреляционным сетям, подход еще не был адаптирован для учета нефиксированной временной динамики. Несмотря на это, для ERGM были разработаны варианты, учитывающие взвешенные ребра, примером которых является обобщенная ERGM (GERGM), и расширение этих вариантов для учета временной динамики представляется несложной задачей. Поскольку в фокусе внимания TERGM моделирование ребер, подход рассматривает агентность акторов лишь в отчасти. Можно построить TERGM вообще без включения атрибутов вершин.

Главным отличием SAOM является их явный акцент на роли акторов. В них определение исходящих связей рассматривается как неотъемлемая часть *поведения* актора. В рамках SAOM существуют два различных процесса, связанных с временем и характером модификации сетей. Эти процессы специально разработаны для отражения принципа самостоятельности акторов [369]. Другими словами, стохастические процессы предполагают, что акторы обладают как способностью, так и мотивацией изменять свои сетевые связи.

Важно уточнить, что “акторно-ориентированная” специфика SAOM не означает, что модель в первую очередь ориентирована на переменные на уровне акторов. Вместе с этим, как и в TERGM, SAOM обращает внимание на ребра и топологию сети, возникающую на их основе. По сути, SAOM представляет собой структуру для понимания изменений во взаимоотношениях между акторами, тем самым изображая модель эволюции сети, а не модель, сосредоточенную на самих акторах.

### 3.4 Базовая модель SAOM

Теперь перейдем к объяснению базовой работы и теоретических оснований, которые отличают тип моделей SAOM. Впервые предложенные Т. Снайдерсом, акторно-ориентированные стохастические модели направлены на интеграцию между методологическими и теоретическими требованиями анализа эволюции сетей [[362]][364]. Базовая модель должна содержать как минимум два наблюдения и одинаковый набор вершин. Предполагается, что такая сеть со стабильным набором акторов изменяется в каждой точке времени потому, что (а) каждый автор знает общую структуру сети и свою позицию в ней, и (б) пытается достичь своих целей в виде наиболее эффективной позиции с помощью изменений отношений с другими акторами [364].

Такой “методологический индивидуализм” ограничивается социальной средой и ресурсами акторов, поэтому модель содержит эвристику приблизительной возможной полезности в каждый момент времени для каждого актора. Эволюция сети определяется, с одной стороны, как объективная функция решений участников, и с другой - как стохастическая функция случайного эффекта, добавляемая в модель для объяснения результатов, не включенных в теорию.

Модель рассматривается как непрерывная марковская цепь [290], предполагающая условную независимость между диадами. Модель преодолевает некоторые недостатки предыдущих моделей, такие как условно однородные модели [173], которые не подходили для проверки широкого круга социологических теорий. Подход SAOM позволяет оценивать не только параметры, описывающие эффекты ковариат индивидуального уровня, но и ковариаты сетевого уровня и сетевых зависимостей [62]. В любой момент времени может быть изменена только одна связь и акторы не координируют свои действия между собой [368]. Более того, в модель заложена миопия, то есть связи формируются только исходя из краткосрочных целей акторов [364].

Эмпирическая работа с моделью может быть проведена благодаря бесплатному пакету в *R*, который называется *RSiena* (Simulation Investigation for Empirical Network Analysis) [325].

Для ее использования необходимы панельные сетевые данные. Основные функции состоят в следующем [363]:

1. функции для спецификации объектов данных (такие как ковариаты, зависимые переменные, и т.д.);
2. функции для создания объектов эффектов модели, обозначенные как *sienaEffects*;
3. функции для подгонки параметров модели: *siena07*, *sienaBayes*, *siena08*;
4. функция для анализа качества модели: *sienaTimeTest*, *sienaGOF* (“goodness of fit”);
5. различные функции для описания результатов.

В модель также добавлены различные диаграммы и переменные для верификации пространства состояний [363]. Эти переменные часто обозначаются как “поведение” и помогают представить коэволюцию между сетями и поведением модели. Так, в классической SAOM, существует одна зависимая переменная - направленная сеть, - а в расширенных версиях могут также присутствовать несколько зависимых сетей.

### 3.5 Оптимизация и оценка

Для оценки SAOM используются различные статистические методы, наиболее распространенным из которых является метод моментов (МоМ) или метод максимального правдоподобия (MLE).

Метод моментов (МоМ) – это метод имитационного моделирования, используемый для оценки параметров в SAOM, подробно описанный Снайдерсом [364]. Он заключается в сравнении описательных статистик наблюдаемой сети со значениями, полученными в результате моделирования при разных значениях параметров. Целью является обнаружение значений гиперпараметров, минимизирующих разницу между наблюдаемой и моделируемой статистикой сети.

Сначала оценки параметров часто задаются произвольно, а затем, путем итерационного моделирования SAOM с различными наборами значений параметров, рассчитываются сводные статистики и сравниваются с соответствующими характеристиками наблюдаемых данных на основе функции расхождения, которая количественно оценивает разницу между наблюдаемой и моделируемой статистиками. Затем оценки параметров обновляются таким образом, чтобы минимизировать функцию расхождения. Функции расхождения, используемые в SAOM, могут оценивать такие свойства, как количество связей, транзитивность, распределение мер центральности, паттерны образования и распада связей, вклад характеристик акторов, специфические параметры диад и временная динамика. Выбор функции расхождения зависит от вопроса исследования, конкретной используемой SAOM и характеристик наблюдаемых сетевых данных.

Несмотря на концептуальную простоту МоМ и возможность работы со сложными SAOM, а также гибкую спецификацию модели, подходящую для различных типов сетевых данных, ее использование сопряжено с определенными трудностями. Для больших сетей или сложных моделей МоМ может быть вычислительно трудоемким, а также требует тщательной настройки алгоритмов оптимизации. Предпринимаются попытки повысить эффективность и расширить спектр использования МоМ. Например, развитие обобщенного метода моментов (GMoM) позволяет обогатить оценку временными данными, вводя в нее в качестве параметров статистику из различных временных моментов [25]. Однако этот метод, как предполагают авторы, не может стабильно превосходить традиционный МоМ, в частности, из-за избыточности признаков, что препятствуют сходимости.

При оценке по методу максимального правдоподобия (MLE) необходимо найти такие значения параметров, при которых наблюдаемые данные наиболее вероятны в рамках данной модели. Параметры обновляются таким образом, чтобы максимизировать логарифм функции правдоподобия, выбранной исходя из характера сетевых данных и предположений об их эволюции. Функции правдоподобия разнообразны и подходят для различных типов данных (бинарных, непрерывных, мультиномиальных, событийных и т.д.). MLE дает оценки, которые асимптотически эффективны: увеличение размера выборки связано с ростом точности и уменьшением погрешности. Этот метод широко используется в статистике и может работать с различными спецификациями SAOM. Однако MLE может требовать больших вычислительных затрат, особенно для сложных SAOM. Сходимость к глобальному максимуму функции правдоподобия может быть не гарантирована, а процесс оценки может потребовать тщательной инициализации.

Более того, определение полной функции правдоподобия может стать сложной задачей из-за зависимостей между диадами в процессе эволюции сети. Другими словами, одна диада может влиять на

поведение других диад в сети. Дальнейшая адаптация MLE – оценка максимального псевдоправдоподобия (MPLE) – решает эту проблему путем максимизации функции псевдоправдоподобия: вероятность для каждой диады вычисляется на основе наблюдаемого состояния этой диады и состояний соседних диад, заданных SAOM. MPLE менее требователен к вычислениям по сравнению с заданием полного совместного правдоподобия для сложных сетей, поскольку требует моделирования только условных связей между диадами, однако, несмотря на вычислительные преимущества, он не всегда может давать асимптотически эффективные оценки. Бесаг утверждает, что максимальная оценка псевдовероятности отражает “локальную” (пространственную) информацию о соседях, в отличие от оценки максимального правдоподобия, которая отражает “глобальную” информацию о соседях [47]. Более того, Снайдерс утверждает, что результаты исследований показывают, что обычно используемые модели случайных графов имеют скорее глобальную, чем локальную структуру, что в конечном итоге приводит к плохим статистическим свойствам MPLE-оценок [364]. Далее он предполагает, что адаптация спецификации модели, например, подходы, основанные на соседстве, с ограничениями на возможные связи между соседями [301], подходы, основанные на латентном пространстве [[291]][167][347], обладают большими возможностями для решения этой проблемы.

Дальнейшая валидация модели, а также сравнение SAOM с различными характеристиками осуществляется с помощью таких тестов качества, как:

1. *Goodness-of-fit (GOF)* тесты оценивают, насколько хорошо SAOM воспроизводит наблюдаемые сетевые данные, сравнивая статистики сетей. Тесты GOF могут использовать имитационное тестирование или методы бутстрепа, но страдают от переобучения, плохой генерализации и чувствительности к размеру выборки. Lospinoso и Snijders [246] предлагают в качестве решения этой проблемы вспомогательные статистики (например, характеристики триад, транзитивность), не включенные в модель в явном виде. Они моделируют расстояние Махalanобиса между вектором вспомогательной статистики и оценкой модели с помощью симуляций Монте-Карло, повторно используя их из вычислений МОМ в SAOM. Вводя собственный принцип минимального описания модели (MMD), они анализируют влияние вспомогательных статистик на GOF, добиваясь баланса между сложностью модели и ее описательной способностью.

2. *Критерии отбора моделей*, такие как информационный критерий Акаике (AIC) или Байесовский информационный критерий (BIC), предлагают количественную сравнительную меру для SAOM: чем меньше значения, тем лучше модель подходит под данные. AIC совмещает оценку соответствия модели данным и штраф за сложность модели:

$$AIC = -2 * \log(\text{likelihood}) + 2 * \text{number of model parameters}$$

BIC штрафует сложность модель сильнее, чем AIC. Этот критерий рассчитывается как:

$$BIC = -2 * \log(\text{likelihood}) + \log(\text{samplesize}) * \text{number of model parameters} *$$

3. *Тесты на сходимость* позволяют определить, сходится ли алгоритм оценки, используемый для подгонки SAOM, к стабильным оценкам параметров. Визуальное изучение графиков параметров модели может помочь выявить проблемы сходимости. В идеале графики должны стабилизироваться по мере выполнения оценки. В отношении несопшедшихся моделей от интерпретации следует отказаться.

Некоторые важные моменты, требующие внимания:

- t-ratio является количественной мерой степени отклонения смоделированной статистики от целевой в среднем.
- Чем меньше t-ratio, тем лучше сходимость. Как правило, t-ratio менее 0,1 считается показателем хорошей сходимости.
- Чтобы считать модель сходящейся, общее максимальное t-ratio сходимости не должно превышать 0,25.
- В тех случаях, когда модель не сходится, рекомендуется повторно провести анализ с использованием опции “prevAns”.

### 3.6 Вырождение

Другой проблемой, возникающей при оценке SAOM, является вырождение. В работах Штрауса [377], Снайдерса [365] и Хэндкока [153] показано, что экспоненциальные модели случайных графов могут быть почти вырожденными, и то же самое может иметь место для SAOM в перспективе отсутствующих временных лимитов (хотя на практике время обычно ограничено). Вырожденность в SAOM возникает, когда несколько наборов значений параметров приводят к одной и той же наблюдаемой структуре сети. Это может затруднить оценку “истинных” или наиболее точных значений параметров и точное определение механизма, управляющего эволюцией социальной сети. Проблема вырождения представляется особенно опасной в сетевом анализе, поскольку сходимость к целевому распределению становится еще более медленной и менее устойчивой в мультиodalных сетях, где типичные алгоритмы, обновляющие отдельные связи или структурные элементы, имеют ничтожно малую вероятность перемещения между модальными областями [365].

Для решения проблемы вырождения SAOM исследователи обычно используют различные стратегии [153], такие как проверка робастности, сравнение различных инициализаций модели, предоставление дополнительных данных для обучения.

В этих практиках также отдается предпочтение байесовскому фреймворку [291]. Помимо уменьшения вырождения модели, он облегчает распространение неопределенности параметров на окончательный вывод и позволяет учитывать предварительные знания экспертов, если они существуют [153]. Кроме того, Лоспиносо и др. [246] предполагают, что введение в модель временной неоднородности может снять проблему вырождения. Временная неоднородность добавляет временное измерение в модель, делая ее более способной улавливать и различать различные состояния сети в разные моменты времени, что, в свою очередь, приводит к улучшению предсказательной силы и качества подгонки, а также позволяет вводить временные ограничения и включать внешние события в качестве параметров модели. Проблема вырождения в бимодальных сетях может быть минимизирована путем адаптации методов оценки, как это было предложено в работе [365].

### 3.7 Заключение

Работа позволила выявить отличительные особенности SAOM как особого подхода к пониманию сетевой динамики, социального влияния и взаимосвязей в многоуровневых сетях. SAOM предлагают

универсальную и всеобъемлющую методологию для исследования сложных сетевых явлений.

В последнее время SAOM доказали свою универсальность в различных областях: от анализа небольших сетей дружбы подростков до политологического анализа транснациональных союзов. Однако их теоретические ограничения и требования к данным накладывают достаточно серьезные ограничения. Современные исследования направлены на решение этих проблем, в то время как применение SAOM для анализа российского общества остается практически неизученным. Благодаря возможностям, открывающимся в различных областях, и текущим проектам, SAOMs открывают перспективы для более глубокого понимания динамики сложных сетей.

Проведен сравнительный анализ SAOM и TERGM, который позволил выявить сильные стороны и уникальный вклад SAOM в изучение динамики сетей. Это сравнение служит ценным ориентиром для исследователей при выборе наиболее подходящей методологии для решения конкретных исследовательских задач. “Акторно-ориентированная” специфика SAOM не означает, что модель в первую очередь ориентирована на переменные на уровне акторов. Напротив, как и TERGM, SAOM обращает внимание на ребра и топологию сети, возникающую на их основе. По сути, SAOM представляет собой структуру для понимания изменений в отношениях между акторами, тем самым изображая модель эволюции сети, а не модель, сосредоточенную на самих акторах. В отличие от этого, TERGM представляют собой модели, сосредоточенные на ребрах сети.

В работе были рассмотрены алгоритмы, используемые для оценки качества и адекватности моделей, полученных с помощью SAOM, с акцентом на статистические показатели, такие как Goodness of Fit. Эти оценки имеют решающее значение для обеспечения точности и надежности анализа на основе SAOM.

## 3.8 Новые типы сетей и возможности их анализа: Многосторонние сети

### 3.8.1 Введение

По следам конференции IFCS 2022, на которой итальянские коллеги представили оригинальные эмпирические данные о мобильности итальянских студентов, В. Батагель вновь заинтересовался темой многосторонних сетей. Структура их базы данных выглядела следующим образом:

Raw data						Factors					
e	prov	univ	prog	year	w	e	prov	univ	prog	year	w
1	AG	Bicocca	BAL	2008	4	1	1	1	4	1	4
2	AG	Bicocca	Edu	2008	1	2	1	1	5	1	1
3	AG	Bicocca	NsMS	2008	1	3	1	1	9	1	1
4	AG	Bicocca	SsJI	2008	1	4	1	1	11	1	1
5	AG	Bocconi	BAL	2008	11	5	1	2	4	1	11
6	AG	Foscari	AH	2008	1	6	1	3	3	1	1
...						...					
37200	VV	uTorino	AFFV	2017	3	37200	107	72	2	4	3
37201	VV	uTorino	AH	2017	11	37201	107	72	3	4	11
37202	VV	uTorino	BAL	2017	3	37202	107	72	4	4	3
37203	VV	uTorino	Edu	2017	1	37203	107	72	5	4	1
37204	VV	uTorino	ICT	2017	1	37204	107	72	8	4	1
37205	VV	uTorino	SsJI	2017	3	37205	107	72	11	4	3

Figure 1: Структура данных для построения многосторонней сети

В данном случае, *стороны* - это векторы (переменные), такие как провинция (prov), университет

(univ), образовательная программа (prog) и год (year). Путем соответствуют веса (w) - сколько студентов были попали на мобильность в эти провинции, университет, программу и год. Такая структура данных называется *взвешенной многосторонней (multiway) сетью*.

Стратегия анализа этих данных со стороны моих коллег заключалась в следующем. Они разделили данные по каждому году отдельно (т.к. уникальных значений было немного, это было легко), и в каждом из этих наборов данных соединили пути “университет” и “образовательная программа” в один. Таким образом, у них получилась взвешенная двудольная (two-mode, bipartite) сеть (граф), методы анализа которых уже давно известны. В. Батагель трансформировал эти данные в R и создал их визуализацию, которую сохранил в формате json. Также он добавил дополнительные региональные данные из открытых источников.

### 3.8.2 Многосторонние сети

**3.8.2.1 Формальное определение и свойства** Взвешенная многосторонняя сеть  $N = (V, L, w)$  состоит из узлов конечного числа множеств  $k$  (сторон, измерений)  $V = (V_1, V_2, \dots, V_k)$ , множества ребер  $L$  и весов  $w : L \rightarrow \mathbb{R}$ . Множества узлов и ребер соединяются с помощью функции инцидентности  $I : L \rightarrow V_1 \times V_2 \times \dots \times V_k$ , которая назначает для каждого ребра  $e \in L$  его  $k$ -кортеж из узлов  $I(e) = [e(1), e(2), \dots, e(i), \dots, e(k)], e(i) \in V_1$ . Если для некоторых разных узлов  $i \neq j$  совпадают их измерения  $V_i = V_j$ , мы говорим, что множества  $V_i, V_j$  имеют один порядок (mode).

В многосторонней сети как для набора узлов, так и ребер могут быть известны дополнительные атрибуты  $N = (V, L, P, W)$ , где  $P$  - это множество узловых атрибутов  $p : V_1 \rightarrow S_p$  а  $W$  - множество реберных весов  $w : L \rightarrow S_w$

Для подмножества ребер  $L' \subseteq L$  мы обозначаем

$$V_i(L') = \{e(i) : e \in L'\}$$

множество узлов стороны  $V_i$  которые являются конечными точками ребра, исходящего из  $L'$ .

Многосторонняя сеть  $N$  называется *простой* тогда и только тогда, когда все ее реберные кортежи взаимно различны:  $\forall e, f \in L, e \neq f \rightarrow I(e) \neq I(f)$ .

Многосторонняя сеть  $N' = (V', L', P', W')$  называется *подсетью* многосторонней сети  $N = (V, L, P, W)$  тогда и только тогда, когда ее можно получить из  $N$  за счет удаления: - некоторых узлов. Если мы убираем узел  $v \in V_i$  мы также убираем все ребра  $e \in L$ , которые содержат в себе этот узел  $e(i) = v$  - некоторых ребер - некоторых сторон - некоторых атрибутов из множеств  $P$  или  $W$ .

Для многосторонних сетей также возможно говорить об их гомоморфизме/изоморфизме.

*Звездой* узла  $u \in V_i$  называется множество ребер такое, что

$$S(u) = \{e \in L : e(i) = u\}.$$

**3.8.2.2 Многосторонний анализ** Помимо понятия многосторонних сетей, в области анализа данных также существует подход многостороннего анализа (multiway analysis) [220]. Он опирается на схожее представление о данных как многосторонних структурах. Однако в отличие от сетевого анализа, многосторонний анализ находится в русле традиционного многофакторного статистического анализа,

который опирается на линейную алгебру, а не реляционную алгебру и теорию графов как сетевой анализ [278], который можно назвать “комбинаторным подходом”.

**3.8.2.3 Опыт реализации** В. Батагель уже работал над косвенным подходом к блокмоделингу дихотомизированных 3-сторонних сетей [40]. В частности, с коллегами мы разработали подход к блокмоделингу 3-сторонних сетей. Визуализация его работы - на рисунке ниже.

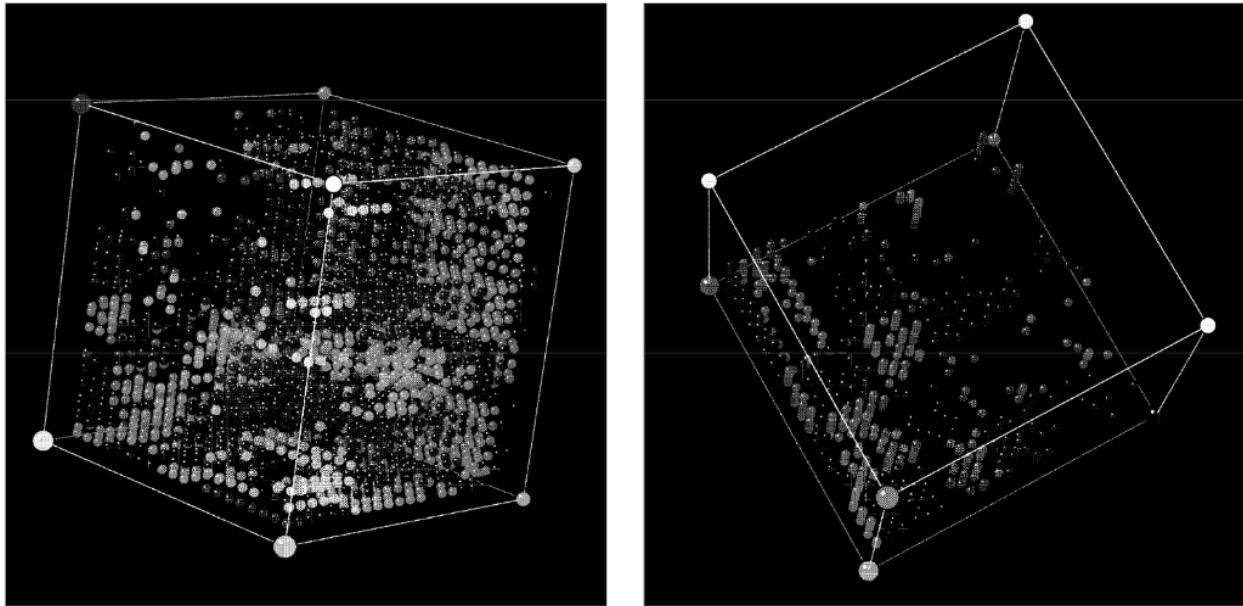


Figure 2: Блокмоделинг 3-сторонней сети

Для схожих целей он позже разработал целый пакет в R [[batagelj2023?](#)].

### 3.8.3 Кластеризация

### 3.8.4 Кейс: Европейские аэропорты

### 3.8.5 Ядра

### 3.8.6 Кейс: Олимпийские игры

## 3.9 Применение современных методов машинного обучения для предсказания связей в социальных сетях

С каждым годом компьютерные технологии все глубже интегрируются в различные научные дисциплины. Использование искусственного интеллекта, нейронных сетей в таких далеких от математики областях, как психология, филология, литературоведение, растениеводство и т.д. становится обыденностью. В данном параграфе описаны возможности использования контролируемого (supervised) и неконтролируемого (unsupervised) машинного обучения (ML) для предсказания связей (link prediction) в социальных сетях.

В разделе представлен сравнительный анализ литературы в парадигме экспертной методологии, которая предполагает качественную стратегию экспериментального отбора источников на основе анализа литературы по теме исследования, что позволяет преодолеть барьеры, связанные с усложнением, специализацией и фрагментацией научных областей, а также с ростом публикационной активности

исследователей, характерным для развития современной науки. Использовалась литература по применению машинного обучения в области социального сетевого анализа (social network analysis – SNA), причем более подробно рассматривалось использование контролируемых и неконтролируемых методов машинного обучения для прогнозирования связей в социальных сетях.

В начале каждого исследования ученым необходимо собрать данные для своего проекта. Но редко данные поступают в виде удобном для обработки. Часто они не структурированы, загрязнены шумом и ненужной информацией. А если сложных данных слишком много, то на их ручную обработку уходит много ресурсов и времени, в течение которого исследование может стать неактуальным. Поэтому уже на этом этапе машинное обучение очень помогает ученым и значительно ускоряет их работу.

Рассмотрим пример, когда нейронные сети предварительно обработали данные в исследовании [159]. Ученые собрали 34672 твита с 1 по 20 апреля 2020 года с необходимыми темами и ключевыми словами. Для дальнейшего анализа исследователям необходимо было отсортировать сообщения, относящиеся к теме исследования, и те, которые просто содержали подходящий набор слов, а затем определить, какие из отсортированных сообщений отражают позицию противников движения, а какие – последователей. Для этого они создали программу, использующую неконтролируемый подход к машинному обучению, в котором применяются тематическое моделирование и обработка естественного языка (natural language processing – NLP). Эта технология предназначена для выявления закономерностей в данных и обобщения содержания твитов в отдельные темы с высокой степенью корреляции. Исследователи использовали модель Biterm Topic Model (BTM), которая выявляет закономерности в коротких текстах. Этот метод кластеризации тем моделирует совпадение слов, что повышает производительность для документов с небольшим разреженным текстом, таких как твиты. После проведенных манипуляций ученые получили актуальные данные, рассортированные по кластерам, что позволило им в дальнейшем провести сетевой анализ и обнаружить интересные закономерности структурирования и влияния общественного мнения.

В нашем мире существует множество вещей и аспектов жизни, по которым людей можно разделить на условные группы, например: интересы и увлечения, сообщества, взгляды и принадлежность. Таким образом, каждый человек в социальной сети может быть охарактеризован набором меток. Однако в реальной работе маркировка занимает много времени и является дорогостоящей, поэтому люди маркируются либо частично, либо достаточно редко. Задача классификации узлов состоит в том, чтобы с учетом структуры сети предсказать метки немаркированных узлов, используя их связи с меченными узлами. Как утверждают Тан и соавторы [382]: «Существующие методы можно разделить на две категории, например, методы, основанные на случайному прохождении, и методы, основанные на извлечении признаков». Метод random walk направлен на распространение меток, а механизм второго метода – на извлечение характеристик узла с использованием информации и статистики, окружающей его.

Ранее работа по классификации сетей строилась следующим образом: сначала извлекались характеристики узлов сети с помощью методов обучения представлению, а затем использовались классификаторы машинного обучения (например, машина опорных векторов, наивный байесовский алгоритмический классификатор и логистическая регрессия для прогнозирования). Сейчас ученые отходят от разделения этапов и разрабатывают структуру, позволяющую объединить эти две задачи таким образом, чтобы отличительная информация, полученная из меток, способствовала обучению

встраиванию сети.

Кластеризация узлов подразумевает разбиение сети на кластеры или подграфы таким образом, что узлы одного кластера более похожи друг на друга, чем узлы других кластеров. В социальных сетях кластеры можно широко наблюдать в виде групп людей с общими интересами или общих сообществ. Ранее основные работы по кластеризации были направлены на кластеризацию сетей с различными показателями близости или силы связи между узлами. Например, на минимизацию количества связей между кластерами с учетом максимизации количества связей внутри кластера. В настоящее время ученые пытаются использовать методы представления сетей для кластеризации узлов. Из этих методов можно выделить те, в которых Tan и соавторы [382] рассматривают «встраивание и кластеризацию как несвязанные задачи, где они сначала встраивают узлы в низкоразмерные векторы, а затем применяют традиционные алгоритмы кластеризации для создания кластеров».

Прогнозирование событий в структурах социальных сетей остается важной исследовательской задачей для SNA. Как утверждает Molokwu [275], «это предполагает понимание внутренних закономерностей связей, сохраняющих заданную структуру социальной сети, на основе изучения ряда структурных свойств, вычисляемых для составляющих ее социальных единиц в пространстве и времени». Часто проблема прогнозирования осложняется тем, что данные о действиях узлов социальной сети скучны или недостаточны.

Еще одним применением машинного обучения в SNA является технология Trend and Pattern Analysis. Эта технология представляет собой модель, обученную на проверенных данных и применяемую к целевым данным. Она позволяет отслеживать и прогнозировать различные результаты действий. Особенно широкое распространение этот метод получил во время пандемии COVID-19. Многие группы ученых изучали это явление с разных сторон, в том числе с помощью Trend and Pattern Analysis. Например, анализ последствий пандемии в нескольких канадских штатах позволил выявить закономерность и получить возможность прогнозировать потребление средств индивидуальной защиты и спрос на них в других географических точках.

В социальных сетях часто встречается недостающая информация: между людьми (узлами) в сети нет связей несмотря на то, что они существуют в реальной жизни. Такие сети являются неполными. Предсказание связей позволяет на основе имеющихся данных об эволюции и структуре сети сделать выводы о ее дальнейшей динамике, а также предсказать будущие связи между узлами. Эта задача очень популярна в настоящее время. Поэтому существует множество способов решения этой задачи с помощью машинного обучения:

- предсказание связей с помощью подхода Strength of Ties,
- предсказание связей с помощью подхода Graph Embeddings,
- предсказание связей с помощью подхода Graph Embeddings на основе матричной факторизации,
- предсказание связей с помощью подхода Graph Embeddings на основе Random Walk(s),
- предсказание связей с помощью подхода Graph Embeddings на основе нейронных сетей [275].

В современном мире процесс установления отношений между социальными субъектами глубоко укоренен в существующих социальных сетях. Социальные сети стали движущим фактором изменения

способа построения социальных взаимодействий, позволяя ускорить создание связей между социальными акторами и сделать поток информации практически безграничным. Вовлеченность социальных субъектов в различные коммуникативные процессы несет в себе ценные данные, которые могут быть использованы для достижения целей в самых разных сферах. Использование зависит от свойств связей между акторами – их прочности, взаимности, возможности будущих связей [407] и т.д.

Например, некоторые интернет-сервисы используют алгоритмы рекомендательных систем, основанные на взаимодействии пользователя с объектом и пользователя с пользователем, чтобы улучшить пользовательский опыт и предложить лучшие контентные решения [183]. Эффективное применение анализа сетевых связей также часто встречается в научных работах, посвященных сетям соавторства [73, 238], в результате чего эта область является одной из наиболее процветающих в последние годы, поскольку способствует развитию эффективной системы совместной работы [77, 183]. К настоящему времени как в научных, так и в практических кругах предсказание связей рассматривается как перспективная область исследований, поскольку несет в себе огромную многоцелевую ценность.

Предсказание связей – это область исследований, которая занимается вопросами прогнозирования социального поведения акторов в социальных сетях [93]. Социальные сети – это динамические образования [228], которые развиваются и изменяются с течением времени, при этом связи исчезают и возникают в силу определенных свойств, связанных с узлами (акторами), структурой группы и т.д. [158]. В предыдущие годы появилось множество научных работ, посвященных острым вопросам предсказания связей, касающихся метрик, используемых для целей предсказания [73, 271], обсуждающих задачу предсказания в различных типах сетей [147, 284]. Наконец, появились обзоры [93, 158, 407], в которых рассматривается вопрос о различных подходах к предсказанию связей. Цель данной части обзора – погрузиться в различные подходы к предсказанию связей и сосредоточиться в первую очередь на алгоритмических подходах машинного обучения. Мотивация данного обзора кроется в стремительном развитии технологий и социальных сетей. Мы также воспользуемся таксономией, предложенной для лучшей систематизации [93], которая помогает структурировать обзор.

Одной из наиболее простых в применении и традиционных групп подходов являются подходы, основанные на сходстве [93, 276]. Методы, основанные на сходстве, исследуют структурную эквивалентность пары узлов, которая затем используется для оценки вероятности будущих связей между этой парой. Высокая степень сходства, соответственно, приводит к повышению вероятности возникновения связей в будущем. При расчетах на основе сходства используются в основном две точки зрения, касающиеся структурного уровня анализа. В подходах, основанных на сходстве, принято использовать локальные и глобальные индексы, а также их современную модификацию, называемую квазилокальными индексами. В основном разница заключается в расстоянии пути до ближайшего узла: при подходе с использованием локальных индексов узлы считаются соседними, если расстояние пути меньше двух [93, 248], а при подходе с использованием глобальных индексов, наоборот, интересны случаи, когда расстояние пути больше двух, что является необходимым условием для того, чтобы узел считался соседним. Квазилокальные подходы, напротив, используют дополнительную топологическую информацию, но в большей степени на локальном уровне [93, 241], учитывают больше информации о соседних узлах [248] и предсказывают другие возможности для связей. Квазилокальный подход развивается и по сей день, при этом вносятся изменения и улучшения в вычислительный алгоритм и метрики, которые часто являются модификациями традиционных метрик локальных индексов [241, 294,

Часто для экономии вычислительного времени и эффективности используются методы, основанные на сходстве, и в этом случае индексы локального сходства оказываются как нельзя кстати, поскольку позволяют одновременно эффективно использовать ресурсы и иметь высокие прогностические характеристики. Однако из-за того, что метрики локального подобия анализируют только пути ближайших соседей, возникает дефицит информации [158], так как упускаются потенциальные связи. Основные работы в области локально-индексных подходов ведутся в области совершенствования метрик вычисления сходства и алгоритмов вычисления сходства [93, 422]. Подходы на основе глобальных индексов используют больше информации и раскрывают больше структуры, однако они крайне неэффективны с точки зрения затрат времени и энергии, поскольку анализируют высокоразмерные связи сетей, что делает их не лучшим выбором для задач предсказания связей. Решение проблемы заключается в снижении размерности и взвешивании сетей, что позволит сократить время вычислений и повысить эффективность прогнозирования [85, 280]. Возможности для совершенствования есть и у квазилокальных подходов, которые сильно зависят от особенностей данных и метрик расчета, выполняемых исследовательской группой. Поэтому существует множество исследований, посвященных устойчивости и робастности квазилокальных инструментов [242, 294, 409]. В целом методы, основанные на подобии, эффективны и зависят от конкретного случая, поэтому исследователям следует внимательно относиться к условиям и задачам исследования и использовать описанные выше подходы.

Следующая группа методов, которые обычно используются для предсказания связей, называется вероятностными. Вероятностные методы используют статистическое моделирование вероятности в соответствии со структурой и размерностью существующей сети. Каждая пара узлов, еще не имеющих связи, включается в модель, которая вычисляет математическую статистическую меру в соответствии с параметрами сети. После вычислений используются гипотезы, которые измеряют степень вероятности того, что эти два узла будут иметь связи в будущем. Такая модель позволяет вписать в предсказание большинство параметров наблюдаемых данных, что делает эту группу подходов более гибкой и универсальной [120]. В этой группе методов исключительно важно обращать внимание на тип сети, так как в марковских и байесовских сетях используются разные методы и расчеты вероятностей [120]. Следует внимательно относиться к типу переменных, взаимности связей, типу взаимных связей и т.д. [212]. Вероятностные модели также имеют возможность работать с множеством измерений и проводить многомерный анализ, однако процесс вычисления структуры и параметров может оказаться непомерно сложным. Используя разграничение, приведенное в работе [93], мы также можем структурировать наш обзор, опираясь на четыре типа подходов в группе вероятностных методов:

1. Модель тензорной факторизации вероятностей [71, 406]. Как указано в [93], они являются логическими расширениями моделей Probability Matrix Factorization, используемых для решения задачи тензорной факторизации. Развитие этого подхода можно увидеть в работах [70, 321, 438].
2. Модель вероятностных латентных переменных [166, 236]. Эти модели развивают идею низкоранговых аппроксимаций для повышения точности предсказания и анализа блоков сетей со схожими свойствами.
3. Марковская модель. Эта группа стохастических моделей показала свою эффективность в применении к динамическим сетям, поскольку позволяет визуализировать эволюцию сети

как процесс [93]. Она также показала более высокую точность предсказания по сравнению с существующими моделями предсказания динамических связей. Из-за большого количества параметров вычисление этой модели занимает много времени, а набор параметров создает дополнительные препятствия для создания модели.

4. Моделирование меток связей [18]. Эта группа моделей применима только к сетям с подписями и решает задачу предсказания меток связей. Подписанные связи несут дополнительную информацию, уникальную для наблюданной связи, но в основном связи можно разделить на две группы по характеру их связи – положительные или отрицательные связи. Положительные связи представляют собой отношения взаимного доверия, близости и общего одобрения. Напротив, негативные связи обозначают антагонистические отношения, характеризующиеся высокой степенью неодобрения и холдности. Это повышает объяснительную силу модели, но одновременно увеличивает время вычислений, поскольку характер связей добавляет в сеть еще одно измерение.

В целом, вероятностные модели обладают более высокой предсказательной точностью и способны нести гораздо больше полезной для анализа информации, однако следует быть осторожным, поскольку дополнительные параметры делают модель более тяжелой и менее устойчивой, поэтому экономия вычислительного времени и ресурсов крайне необходима.

Еще одним подходом, обеспечивающим высокую эффективность прогнозирования, является набор алгоритмических подходов. Они широко представлены в литературе и продолжают развиваться по сей день, поскольку обеспечивают скорость и эффективность. Одним из их преимуществ по сравнению с подходами подобия и вероятностными подходами является возможность использования дополнительной информации из сети, а также использование дополнительной информации, которая может как-то повлиять на формирование связей [93].

Существуют исследования, в которых используются данные о структуре сообществ, что значительно повышает точность предсказания; одной из наиболее востребованных для получения результатов информации являются данные о поведении пользователей, которые могут нести в себе многоцелевую информацию. Чтобы более четко организовать обзор этой группы алгоритмов и сфокусироваться на одном конкретном подходе (а именно на машинном обучении), сначала целесообразно привести современные условия и методы, помимо машинного обучения. Подход машинного обучения будет рассмотрен позже и более подробно. Мы также используем разграничение, приведенное в работе [93].

Существует три группы подходов – метаэвристические, матричной факторизации и машинного обучения. Метаэвристическая группа методов содержит рекомендации и приемы, которые помогут исследователю применить наилучший вариант эвристического метода оптимизации. Эта группа была широко использована в исследовании [93], показав свою высокую эффективность в задачах, где анализировались большие сети, поскольку требовала меньшего времени и вычислительной эффективности по сравнению с другими алгоритмами. Исследования предполагают дальнейшее развитие этих подходов, поскольку они тестируются на большем количестве информации и характеристик сети с целью повышения их точности. Факторизация матриц – группа алгоритмов, использующих колаборативную фильтрацию [93], которая принимает в качестве результата предсказания произведение, образующееся после слияния двух матриц меньшей размерности. Однако эта методика не столь надежна и стабильна, поскольку

зависит от данных – если есть шум, выбросы, смещение или экстремальная дисперсия, то предсказание не будет стабильным и точным. Решить эту проблему достаточно просто, так как данная группа подходов является универсальной и гибкой, поэтому во многих работах уже рекомендуется использовать методы мешков, которые хотя и увеличивают время вычислений, но при этом значительно повышают точность прогнозирования. Данный подход также применяется, когда ставится задача анализа неявных признаков динамической сети. Машинное обучение сочетает в себе достижения предыдущих алгоритмов, а также позволяет сэкономить гораздо больше вычислительного времени и усилий.

Контролируемое обучение – это одна из ветвей машинного обучения, известная как Supervised Machine Learning (SML). Этот метод отличается от других тем, что для обучения алгоритмов классификации данных или точного прогнозирования результатов используются полностью помеченные наборы данных. Наличие полностью маркированного набора данных означает, что каждый пример в обучающем наборе имеет правильный ответ, и цель алгоритма – получить этот ответ. Таким образом, помеченный набор данных с фотографиями фруктов позволит обучить нейронную сеть с фотографиями яблок, груш, бананов и т.д. Когда сеть получает новую фотографию фрукта, она сравнивает ее с примерами из обучающего набора данных, чтобы предсказать ответ.

Существует множество алгоритмов и вычислительных методик контролируемого обучения. Можно выделить несколько часто используемых методов: нейронные сети, Naive Bayes, линейная и логистическая регрессия, support vector machines (SVM), K-nearest neighbor, Random forest.

Контролируемое обучение имеет как преимущества, так и недостатки. К числу преимуществ контролируемого обучения относятся:

- простота обучения: поскольку алгоритм обучается на помеченных данных, обучать модель гораздо проще. Кроме того, этот процесс прост в случае реализации и понимания процессов;
- ясность данных: Каждый алгоритм контролируемого обучения использует помеченные данные, поэтому входные данные должны быть отнесены к определенным категориям, что позволяет уменьшить количество ошибок при работе с этими данными;
- прогнозы, как правило, более точны и надежны, если имеется достаточное количество соответствующих данных.

Несмотря на то, что контролируемое обучение имеет ряд преимуществ, при построении моделей такого типа возникают определенные трудности:

- при работе с большими и сложными наборами данных алгоритмы контролируемого обучения могут быть сравнительно более трудоемкими и вычислительно дорогими;
- при работе с большими наборами данных возрастает вероятность человеческой ошибки, приводящей к неправильному обучению алгоритмов;
- для контролируемого обучения необходимы помеченные данные, то есть данные должны быть классифицированы по определенным категориям, прежде чем алгоритм сможет на них обучаться.

Неконтролируемое обучение как метод автоматизированной обработки данных берет свое начало с перцептрона, построенного в 1958 году Фрэнком Розенблаттом. Перцепtron классифицировал

примитивные изображения, используя солнечные (фотоэлектрические) элементы. Однако, пройдя значительный путь эволюции, сегодня алгоритмы неконтролируемого обучения стали действительно мощным инструментом. Например, неконтролируемое обучение может быть использовано для выравнивания графов знаний или построения вкраплений графов.

Основное преимущество бесконтрольного обучения заключается в том, что обучающий набор данных не обязательно должен быть помечен, т.е. для обучения сети не нужно давать правильные ответы или решения. В случае отсутствия помеченных данных бесподчиненное обучение часто оказывается более дешевым и быстрым решением, чем создание помеченного обучающего набора данных.

Для предсказания связей в социальных сетях обучение без контроля впервые было использовано в 2007 году Либен-Ноуэллом и Клейнбергом. Они изучали временные соавторские сети и использовали граф, соответствующий более раннему состоянию сети, для предсказания новых связей в сети, соответствующих более позднему периоду времени.

Большинство алгоритмов предсказания связей без наблюдения используют сходство между узлами для предсказания того, должна ли быть сформирована связь. Ниже приведены некоторые популярные методы вычисления этого сходства:

- общие соседи (Common Neighbors, CN) - более высокая вероятность предсказания ребер между узлами с большим числом общих соседей;
- алгоритм Jaccard (Jac) - зависит от количества общих и разных соседей у двух узлов;
- алгоритм Лейхта-Холма-Ньюмана (LHN) - сравнивает реальное количество общих соседей с ожидаемым количеством общих соседей;
- алгоритм Адамика-Адара (AA) - также дает более высокую вероятность предсказания ребер между узлами с большим числом общих соседей;
- алгоритм Local Path (LP) - этот алгоритм также учитывает 2- и 3-хоповых соседей.

Статистика является широко используемым инструментом для предсказания связей, например, вероятностная мягкая логика (PSL) [140] и марковские логические сети (MLN) [324]. Неподконтрольный механизм, использующий статистику, был реализован Куо и др [222]. Они работали с социальными онлайновыми сетями и анонимными отзывами пользователей. Вопрос, который задают авторы, заключается в следующем: «можем ли мы предсказать носителя мнения в гетерогенной социальной сети без каких-либо помеченных данных?».

Их алгоритм получил название «Factor Graph Model with Aggregative Statistics (FGM-AS)». В его основе лежат три слоя: “Кандидат”, “Атрибут” и “Счет”. Слой “Кандидат” - это слой с парами случайных вершин, которые потенциально могут иметь общее ребро. Слой “Атрибуты” содержит атрибутивную информацию о кандидатах. Слой “Count” кодирует агрегированную статистику кандидатов. Таким образом, исследователи используют три типа функций: Функции “атрибут-кандидат”, “кандидат-кандидат” и “кандидат-счет”.

Предсказание связей в социальных сетях часто является задачей прогнозирования временных изменений в графе. Соавторские сети, с которых началась история ненаблюдаемого предсказания связей

[238], также изучались Мунисом и др [280] во временной перспективе. Они объединили контекстную, временную и топологическую информацию для предсказания связей в соавторской сети.

Многие из современных механизмов предсказания связей без наблюдения используют ту же идею, что и Либен-Ноуэлл и Клейнберг [238]: использование старых ссылок в качестве обучающего множества и новых ссылок в качестве тестирующего множества. Однако при этом важно, какой тип вкраплений графа используется. Поэтому данная идея может быть очень удобно реализована с помощью CTDNE - Continuous-Time Dynamic Network Embeddings [288]. Этот алгоритм обучает сеть динамически, внедряя временную информацию во вкрапления графов. Это делает его идеальным для предсказания временных связей в социальных сетях.

Одной из ключевых идей CTDNE является временное случайное блуждание: временно близкие ребра имеют больше шансов быть связанными. Нгуен с соавторами [288] сообщают о среднем выигрыше в качестве предсказания связей в темпоральных графах на 11,9% по сравнению с другими алгоритмами встраивания в сеть DeepWalk, Node2Vec и LINE. Однако эти результаты могут быть сомнительными: DeepWalk, Node2Vec и LINE были представлены более чем за 3 года до CTDNE.

В работе «Towards Fast Evaluation of Unsupervised Link Prediction by Random Sampling Unobserved Links» Ванг и соавторы [406] рассматривают проблему оценки предсказания связей без наблюдения. Основная проблема оценки ненаблюдавшегося предсказания связей заключается в том, что ненаблюдавшихся потенциально возможных связей гораздо больше, чем наблюдаемых. Этот дисбаланс создает трудности для оценки, так как «нереально количественно оценить вероятность существования».

В качестве решения они предлагают выбирать для тестирования только некоторые из ненаблюдавшихся ребер, а не тестировать модели на всех связях сети. Этот метод позволяет быстрее оценивать модели и приводит к значительной стабильности. Однако авторы предостерегают читателей от использования этого метода на небольших сетях, поскольку он может быть рискованным и приводить к худшим результатам.

В рамках задачи изучения контролируемого и неконтролируемого обучения для предсказания связей в социальных сетях рассмотрены возможности использования машинного обучения для решения задач в области сетевого анализа. Описаны особенности задачи предсказания образования связей (предсказания связей) в сетевом анализе для понимания и объяснения процессов, управляемых социальными взаимодействиями. Проведено сравнение и выявлены особенности использования контролируемых и неконтролируемых методов машинного обучения для предсказания связей в социальных сетях. Приведены примеры применения контролируемого и неконтролируемого машинного обучения для предсказания связей в социальных сетях в области социальных наук. Машинное обучение может применяться на любом этапе анализа социальных сетей: предварительная обработка и обработка данных, классификация узлов, кластеризация, анализ на основе событий, анализ тенденций и предсказание связей.

### **3.10 Современные подходы в области сетевой кластеризации и блокмоделинга: систематизация и сравнительный анализ**

#### **3.10.1 Введение**

Цель настоящей главы и приведенного в ней аналитического обзора состоит в систематизации различных подходов в области сетевого кластеризаций и блокмоделинга и их сравнительном анализе для оценки эффективности применения при анализе динамических сетей (на примере сетей соавторства, изменяющихся во времени). Достижение этой цели требует решения следующих задач:

1. Рассмотреть особенности применения методологии блокмоделинга для анализа динамических сетей.
2. Проанализировать детерминистический и стохастический подходы и разработанные в них модели блокмоделинга, используемые для анализа динамических сетей.
3. Сравнить модели блокмоделинга в рамках детерминистического и стохастического подходов, используемые для анализа динамических сетей.
4. Проанализировать возможности применения блокмоделинга для изучения сетей соавторства в различных научных дисциплинах, в т.ч. в динамике.

В работе применялся сравнительный анализ и системный подход для проведения систематического обзора литературы в парадигме экспертовой методологии, сформированной в рамках качественных исследований в области истории и социологии науки. Эта методология предполагает качественную стратегию экспертного отбора источников на основе анализа литературы по теме исследования, что позволяет преодолеть барьеры, связанные с усложнением, специализацией и фрагментацией научных направлений, рост публикационной активности исследователей, характерные для развития современной науки. В качестве основы для сравнения различных подходов использовано моделирование методом Монте-Карло.

В этой главе рассматриваются особенности использования методологии блокмоделинга для анализа динамических сетей, проблематика и актуальность которых показаны в части 1. Поскольку одной из основных особенностей анализа динамических сетей является проверка их устойчивости, там же приведены основные сведения о подходах к оценке устойчивости структуры блок-моделей (Rand Index, Рэнд-индекс). В части 2 представлены модели, разработанные на основе детерминированного и стохастического подходов к блокмоделированию для анализа динамических сетей. В части 3 приведены основные результаты сравнения моделей с помощью имитационного моделирования методом Монте-Карло (метод симуляции), в т.ч. модели для исследования ненаправленных сетей в динамике. В части 5 анализируется использование блокмоделинга для изучения ненаправленных сетей соавторства во времени в различных научных дисциплинах.

#### **3.10.2 1. Особенности применения блокмоделинга для анализа сетей в динамике (tempоральных сетей)**

**3.10.2.1 Блокмоделинг как методология кластеризации сетевых данных** Блокмоделинг — это группа методов кластеризации единиц и связей в сети. По определению Дориана и коллег [39], блокмоделинг используется для сведения больших, сложных, потенциально бессвязных сетевых структур к меньшим, понятным сетям, которые легче интерпретировать. Результатом блокмоделинга являются 1)

кластеры эквивалентных единиц в сети и 2) связи внутри и между этими кластерами. Сжав или уменьшив большую сеть, получается сеть меньшего размера, которая называется блокмоделью. *Блокмодель* — это сеть, в которой узлы образуют кластеры из исследуемой сети, которые также называются *позициями* эквивалентных узлов. Другим результатом блокмоделинга является *матрица изображений*, которая представляет связи между и внутри полученных кластеров. Термин блок относится к части матрицы, которая показывает связи *между* двумя кластерами или *внутри* одного кластера. Диагональные блоки обозначают связи между единицами из одного кластера, а вне-диагональные — связи между единицами из разных кластеров.

Рассмотрим пример на Рис. 1, где представлена сеть узлов и связей между ними. Можно заметить, что узлы, окрашенные в синий цвет, имеют очень похожую структуру связей: они не связаны друг с другом, но все они связаны с узлами, окрашенными в красный цвет. Поскольку у них схожая структура связей, окрашенные в синий цвет узлы можно считать эквивалентными. То же самое относится и к узлам, окрашенным в красный цвет, которые также связаны друг с другом. Поскольку узлы синего и красного цветов имеют схожую структуру связей (или эквивалентны), их можно сжать в один узел, что показано на fig. 3 ниже под визуализацией сети. Из исходной сети мы получаем два новых узла, которые теперь являются кластерами узлов, представляющих их группы. Между этими двумя узлами имеются связи: одна из них соединяет красный и синий узлы, что означает, что эти две группы узлов связаны друг с другом, а также имеется петля, которая визуализирует тот факт, что узлы, окрашенные в красный цвет, тесно связаны между собой.

Другой пример представляет собой взвешенную сеть (связям присвоены некоторые значения), визуализированную в матричном виде и в виде графа (fig. 4). Слева представлена исходная сеть в матричном виде и в виде графа. Применив процедуру блокмоделинга, мы получили *разбиение* (partition). На основе этого разбиения мы можем упорядочить узлы таким образом, чтобы узлы, входящие в один кластер, располагались рядом друг с другом, а узлы из разных кластеров - отдельно (в середине). Различные кластеры разделены синими линиями; видно, что значения внутри каждого блока достаточно однородны по значениям силы связей. В исходной сети узлы окрашены в соответствии с решением блокмоделинга. Справа представлена полученная блокмодель: *матрица изображений* и *граф изображений*, состоящий всего из трех узлов (кластеров); плотности блоков показаны в соответствующих ячейках этой матрицы. Наблюдается некая структура “ядро-периферия”: ядро и две сплоченные группы, которые внутренне очень хорошо связаны друг с другом и с ядром.

Два узла объединяются в одинаковые кластеры, если они эквивалентны. Существует несколько определений эквивалентности, наиболее известными из которых являются:

1. Структурная эквивалентность [244], White and Reitz:

- a) Две единицы структурно эквивалентны, если они одинаково связаны друг с другом и со всеми остальными единицами.
- b) Две структурно эквивалентные единицы неразличимы, если удалить их лейблы.
- c) Соответствующие идеальные блоки имеют либо все возможные связи (полные блоки), либо ни одной связи (нулевые блоки).
- d) Наиболее часто используемая на практике эквивалентность.

2. Регулярная эквивалентность(White and Reitz

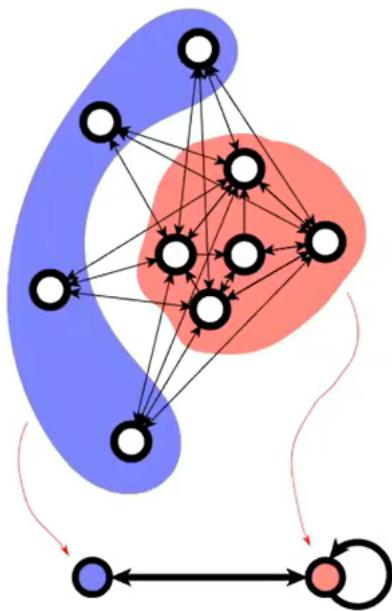


Figure 3: Пример 1: Эмпирические данные (вверху) и блок-модель (внизу) в виде графика

## Example

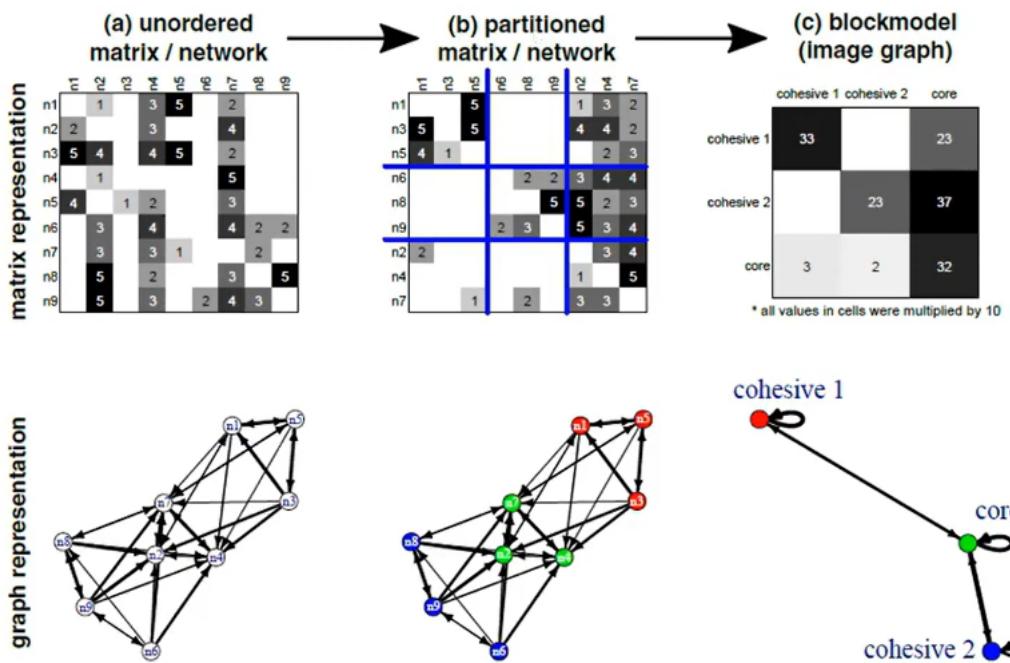


Figure 4: Пример 2: Эмпирические данные в матричном (вверху) и графическом (внизу) представлении.

- a) Регулярно эквивалентные единицы соединяются одинаковым образом с кластерами эквивалентных единиц.
- b) Это определение очень близко к социологическому определению “социальной роли”, однако оно редко встречается в эмпирических данных и (как показали исследования проблемы недостающих данных и устойчивости разбиений) весьма чувствительно к небольшим изменениям в сети.
- c) Соответствующие идеальные блоки имеют хотя бы одну связь в каждой строке и столбце (регулярные блоки) или полностью пусты (нулевые блоки).
- d) Регулярная эквивалентность является частным случаем структурной эквивалентности: если выполняется структурная эквивалентность, то выполняется и регулярная эквивалентность.
- e) Также очень известна.

### 3. Стохастическая эквивалентность Holland et al.

- a) Две единицы статистически эквивалентны, если они имеют одинаковые вероятности связей со всеми другими единицами.

### 4. Обобщенная эквивалентность Doreian et al.

- a) Эквивалентность определяется набором допустимых типов блоков и, возможно, их расположением.
- b) Обобщенный блокмоделинг очень удобен тем, что можно задавать различные типы блоков из множества возможных определенных типов блоков, а это уже косвенно определяет критериальную функцию.

Блокмоделинг - очень универсальный метод, который может быть использован для различных целей в сетевом анализе. Различные типы процедур блокмоделинга могут быть реализованы с помощью специально разработанного пакета “blockmodeling” на языке программирования R. Изначально пакет “blockmodeling” задумывался как реализация обобщенного блокмоделинга для взвешенных сетей. Кроме того, он позволяет вычислять меры сходства и несходства, основанные на структурной и регулярной эквивалентности (алгоритмы REGE), а также строить матрицы разбиения. Это пока единственное программное обеспечение, поддерживающее обобщенный блокмоделинг для взвешенных сетей. Недавно к его функционалу был добавлен анализ обобщенного блокмоделинга связных (например, многоуровневых) сетей.

**3.10.2.2 Динамические сети** Термин “динамические сети” — это общий термин, который охватывает различные типы сетей (например, сети с временной меткой событий или моментальные «снимки»), включающие некоторую временную информацию. Подробное описание различных типов динамических сетей (или темпоральных) см. в работе [177].

В работе мы сосредоточимся на сетях, которые измеряются в нескольких периодах времени, также называемых сетями в формате моментальных «снимков», или snapshot-сетями [260], [447]. Сети snapshot являются одной из разновидностей динамических сетей, в которых большинство узлов присутствуют во всех временных точках и измеряются одни и те же типы отношений. Например, исследование дружеских отношений между старшеклассниками, которое проводилось в трех временных точках - один раз в

феврале, затем в марте и затем в апреле одного года (fig. 5). Во всех временных точках мы измеряем один и тот же тип отношений, которым является дружба. Таким образом, мы получаем три сети, наблюдаемые в каждой временной точке. Мы можем анализировать эти сети отдельно независимо друг от друга, но можем и заметить, что одни и те же единицы, наблюдаемые в разных временных точках, и связи из разных временных точек являются зависимыми.

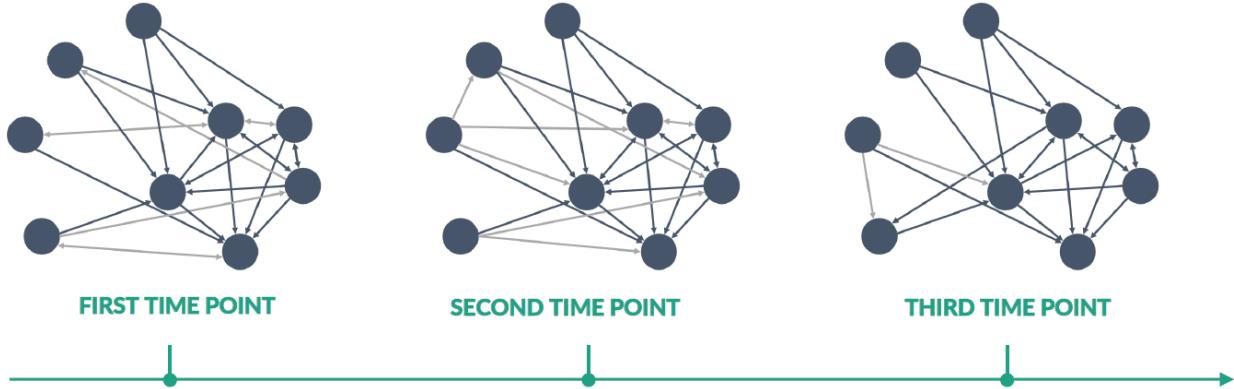


Figure 5: Сеть дружбы среди старшеклассников в 3 временных периодах

Такие сети можно также представить в виде визуализации матричного типа на fig. 6. Мы включаем в это понятие типы сетей, также известные как «последовательности сетей или графов» или «многослойные сети» [177, с. 5] или являющиеся продуктом «представления с потерями» [177, с. 8–9] некоторых других типов, например, взвешенных графов и графов «временного окна», согласно Holme [177, с. 8–9]. Таким образом, моментальные snapshot-сети состоят из серии измерений (не слишком больших) одного и того же отношения между (хотя бы частично) одними и теми же единицами.

Следует отметить, что в данном отчете мы не рассматриваем другие виды сетей, к которым относятся сетевые данные с временной меткой, или повторяющиеся события взаимодействия [84], реляционные события / данные [59] и последовательности контактов [177].

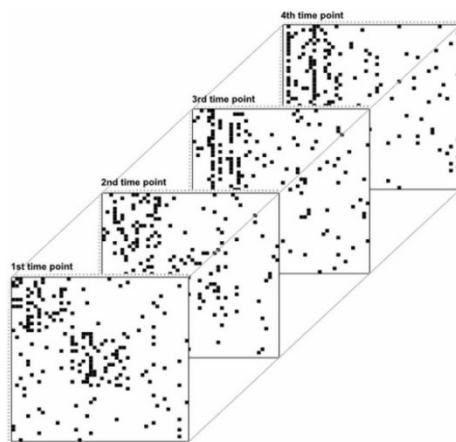


Figure 6: Коллекция сетей в разные моменты времени (динамическая сеть): два измерения представляют единицы (отправители и получатели связей), а третье измерение - моменты времени

**3.10.2.3 Применение блокмоделинга для динамических сетей** Подходы к блокмоделингу сетей, наблюдаемых в один момент времени, имеют давнюю традицию и хорошо изучены и оценены применительно к различным реальным и синтезированным сетям [38, 121, 244, 346, 366, 444, 448].

Только в последние годы исследователи предложили подходы к блокмоделингу динамических сетей. Идея блокмоделинга динамических сетей заключается в использовании того факта, что последовательно наблюдаемые сети являются зависимыми. Исследователи выдвинули идею о том, что учет зависимости между связями из разных временных точек может повысить достоверность результатов. Хотя цель блокмоделинга динамических сетей остается той же, что и при статическом блок-моделировании, - выявление эквивалентных групп единиц и связей между выявленными группами, - при изучении динамических сетей блокмоделинг проводится для сети в каждой временной точке, чтобы изучить, как изменяются структуры во времени.

Целью “обычного” блокмоделинга одномодальных сетей является поиск групп и связей между ними в одной одномодальной сети (т.е. сети со связями между одним набором единиц, измеряемых только в один момент времени). Существует множество подходов к блокмоделингу одномодальных сетей, такие как обобщенный блокмоделинг [39, 443], блокмоделинг методом k-means [57, 156, 447] и стохастический блокмоделинг [135, 148, 303, 366]. Практически все подходы к одномодальному блокмоделингу оценивают разбиение. Эти подходы также оценивают набор параметров или вычисляют статистики, описывающие связи между полученными группами, которые называются блокмоделью.

В отличие от этого, целью блокмоделинга *динамических* сетей является поиск групп и связей между ними в динамической сети для каждого периода времени (*временной точки*) с учетом любой связи между группами из (близких) временных точек. Использование динамических подходов в отличие от «обычного» блокмоделинга для каждой временной точки в отдельности должно дать то преимущество, что при блокмоделировании определенной временной точки информация из других временных точек улучшает результат (в основном разбиение) для этой временной точки. Кроме того, динамическая модель может объяснить или показать, как сеть (структура, разбиение) изменяется во времени. Подходы к блокмоделингу динамических сетей также оценивают разбиения (обычно для каждой временной точки отдельно) и оценивают блокмодель для каждой временной точки. Кроме того, в них обычно каким-либо образом связываются (“параметры” или “решения для”) различных временных точек. Часто они также накладывают определенное ограничение на решение или параметры, например, такое:

- Членство в группе не изменяется во времени,
- Параметры связности или блокмодели (параметры, управляющие/описывающие модель связности между группами) не меняются во времени или существуют некоторые ограничения на эти изменения,
- Все блоки должны присутствовать во всех временных точках.

Большинство подходов к блокмоделингу динамических сетей появилось недавно, и они разрабатывались независимо друг от друга. В этом случае возникает необходимость описать различные подходы к блокмоделингу динамических сетей, сравнить их и дать некоторые рекомендации по выбору подходов к блокмоделингу в различных условиях. Предлагаемые подходы к анализу динамических сетей можно отнести к детерминированному и стохастическому блокмоделингу, поэтому вначале мы опишем и сравним их (глава 2). Затем мы проанализируем подходы к блокмоделингу, которые были разработаны для изучения ненаправленных динамических сетей, например сетей научного сотрудничества (глава 3).

Таким образом, в данной работе мы рассматриваем следующие подходы к блокмоделингу динамических сетей: 1) подходы, в которых осуществляется блокмоделинг сетей- моментальных снимков; 2) подходы, в которых членство в группе и блокмодель могут изменяться во времени. Сюда относятся

как специально разработанные подходы к динамическим сетям, так и подходы, которые могут быть использованы для изучения моментальных сетей, например, подходы к блокмоделированию связанных или многосторонних сетей [35, 357, 446, 447].

В недавней работе [90] оценивались различные подходы к блокмоделингу динамических сетей на имитационных сетях. Сети моделировались методом Монте-Карло таким образом, чтобы они были ближе к реальным сетям, чем сети, созданные на основе простой стохастической блокмодели. Последнее было достигнуто за счет использования подхода к генерации сетей, позволяющего учитывать локальные сетевые механизмы при создании связей внутри блоков. В рамках данного имитационного исследования при сравнении сетей были рассмотрены некоторые вопросы. Во-первых, создание сетей с различными свойствами, чтобы учесть некоторые факторы, которые могут влиять на эффективность подходов блокмоделинга (рассматривались различные характеристики, такие как размер сети, тип блокмодели, устойчивость разделов). Второй момент заключался в том, чтобы генерировать эти сети таким образом, чтобы они были похожи на сети реального мира, что повысит достоверность имитационного исследования. Третья проблема заключалась в том, чтобы учесть некоторые локальные сетевые механизмы при генерации связей внутри или между группами. Также был предложен алгоритм генерации сетей с заданным типом блокмодели и заданным разбиением для оценки эффективности того или иного подхода к блокмоделингу. Результаты этого имитационного исследования представлены ниже.

Следует, однако, отметить, что для тех типов сетей, которые не рассматриваются в данном отчете (временные сети / повторяющиеся события взаимодействия по [261], реляционные события / данные по [59] или последовательности контактов по [177]), также разработаны специальные подходы блокмоделинга (например, [83, 84, 261, 304], и они могут быть проанализированы в дальнейшем. Существует множество подходов к блокмоделингу динамических сетей в широком смысле, т.е. не ограничивающихся сетями, измеренными в нескольких временных точках (обзор см. в [202, 226], раздел SBM с лонгитюдным моделированием), которые также не анализируются ниже.

**3.10.2.4 Оценка устойчивости блокмоделей** Рассмотрение различных начальных разделений является простым и очевидным решением и может привести к улучшению общей эффективности подходов к блокмоделингу. Для измерения сходства двух разделений разработано множество индексов [24]. Большинство из них предполагают наличие единого набора единиц для классификации, что может оказаться нецелесообразным для решения текущей задачи исследования. В случае изучения устойчивости ядер сети соавторства необходимо учитывать, что во втором временном интервале появляются новые исследователи, а другие покидают сеть. Это означает, что разделения получены не для одного и того же набора единиц. Кроме того, разделение и объединение кластеров оказывает одинаковое (негативное) влияние на значение этих показателей.

Эффективной мерой сходства между двумя кластеризациями данных является индекс Рэнда [318]. Индекс Рэнда — это точность определения принадлежности или непринадлежности связи к кластеру. Скорректированный индекс Рэнда (Adjusted Rand Index, ARI) был определен как форма индекса Рэнда, скорректированная с учетом случайности группировки элементов. Цель скорректированного индекса Рэнда - сравнить сходство между двумя результатами кластеризации. Более высокие значения указывают на большее сходство разделов. При точном совпадении двух разделений значение индекса равно 1. В случае двух случайных разделений ожидаемое значение индекса равно 0. Это позволяет сравнивать

сходство разделений, полученных из разных сетей, разных размеров и с разным количеством кластеров.

Цугмас и Ферлигой [89] предложили три версии модифицированного индекса Рэнда, которые не предполагают единого набора единиц для классификации (один набор является подмножеством другого набора).

*Модифицированный индекс Рэнда 1 (MRI 1)* предполагает, что второй набор единиц для классификации является подмножеством первого набора единиц. Данные первого и второго набора единиц обычно измеряются в двух временных точках (периодах).

На примере сетей соавторства термин “единица” может обозначать исследователей, включенных в блокмоделинг. В контексте измерения устойчивости ядер исследователи, присутствующие в ядрах в первый период времени, но не присутствующие в сети во второй период времени (например, ушедшие на пенсию или не участвующие в соавторстве), или исследователи, которые были в одном из ядер в первый период времени, но находятся на полупериферии или периферии во второй период времени, обозначаются как выбывшие. Исследователи, отсутствовавшие в первом временном периоде (пришедшие), или исследователи, отнесенные в блокмодели к неосновной части, были удалены из сети, так как они не влияют на меру стабильности.

Мера определяется на интервале от 0 до 1, где большее значение меры указывает на более устойчивую классификацию. Значение 1 возможно только при отсутствии выпадающих кластеров. Объединение кластеров во втором временном интервале не приводит к снижению значения меры, в то время как разделение приводит. Предполагается, что объединение двух ядер не должно снижать значение меры устойчивости, поскольку объединение свидетельствует о большем количестве связей (соавторства) между исследователями, а это не влияет на связи, созданные в первом временном периоде. С другой стороны, если бы объединение ядер увеличивало значение меры, то максимальное значение меры было бы  $>1$ .

Поскольку ожидаемое значение двух случайных и независимых разбиений не принимает постоянного значения, Цугмас и Ферлигой (2015) предложили корректировку. Корректировка может быть получена путем моделирования, в котором порядок единиц двух разделений U и V независимо и случайно переставляется много раз. Для каждой перестановки обоих разделов вычисляется значение MRT 1. Среднее из полученных значений представляет собой ожидаемое значение MRI 1 в случае двух случайных и независимых разделений. Значение *модифицированного скорректированного индекса Рэнда 1 (MARI 1)* рассчитывается следующим образом:

$$MARI1 = \frac{MRI1 - E(MRI1)}{1 - E(MRI1)} \quad (1)$$

Ожидаемое значение MARI 1 в случае двух случайных и независимых разделений равно 0, а максимальное - 1. Как уже отмечалось, в случае измерения устойчивости ядер блокмоделей соавторства по MRI 1 необходимо удалять из сети входящих исследователей. В процессе получения ожидаемого значения меры в случае двух случайных и независимых разбиений это можно делать до или после каждой перестановки исследователей каждого разбиения. В случае изучения устойчивости ядер сетей соавторства различия между указанными методами ожидаются очень малы (значение коэффициента корреляции Пирсона на эмпирических данных составляет  $r=0,99$ ). Эти различия могут быть объяснены многими факторами. Одним из наиболее заметных является доля исследователей, отнесенных в первом временном периоде к

ядру, а во втором - к не-ядру.

В случае имитационных исследований (подобных представленному ниже) для сравнения расчетных и истинных разделений можно использовать Скорректированный Индекс Рэнда (ARI). В случае подхода динамического блокмоделинга, для количественной оценки эффективности того или иного подхода к блокмоделингу рассчитывается Скорректированный Индекс Рэнда (ARI) между истинным разделением и разделением, оцененным с помощью предложенных подходов. Этот расчет может быть выполнен для каждой временной точки, и далее мы анализировали среднее значение Индекса Рэнда по всем временным точкам. Скорректированный индекс Рэнда измеряет долю всех возможных пар, которые попадают в один и тот же кластер или в разные кластеры в обоих разделениях. Это позволяет оценить степень согласия между расчетным и истинным разделениями.

### **3.10.3 Детерминированный и стохастический блокмоделинг для анализа динамических сетей**

#### **3.10.3.1 2.1. Детерминированный и стохастический блокмоделинг: общее разделение**

Различные подходы блокмоделинга можно разделить на детерминированные и стохастические.

В детерминированных подходах к блокмоделингу итерационный алгоритм ищет гомогенные блоки по значениям связей. В рамках этого подхода можно выделить 1) *конвенциональный блокмоделинг*, известный также как непрямой блокмоделинг, и 2) *обобщенный блокмоделинг*.

Процедура конвенционального блокмоделинга заключается в вычислении матрицы несходства по матрице смежности сети (можно использовать и другие статистики, описывающие блоки, например, распределение триад), а затем в использовании одного из стандартных алгоритмов кластеризации на этой матрице, например, подходов Ward или K-средних. Минусы такого подхода заключаются в том, что необходимо определить и вычислить совместимую меру сходства. Для эквивалентности структур это исправленное Евклидово расстояние, но иногда может оказаться, что для некоторых типов эквивалентности такой совместимой меры не существует (регулярная эквивалентность - один из таких случаев). Кроме того, не существует явной меры соответствия (fit).

Обобщенный блокмоделинг — это прямой подход, что подразумевает, что оптимизируется эксплицитная критериальная функция, при этом матрица сходства не вычисляется, вся процедура выполняется на уровне сетевых данных. Решение находится путем случайного перемещения узлов в кластере таким образом, чтобы минимизировать значение критериальной функции. В основе лежит обобщенная эквивалентность, то есть эквивалентность определяется на основе допустимых типов блоков. Критериальная функция отражает разницу между идеальной структурой сети и эмпирической структурой сети. Основная процедура вычисления критериальной функции состоит в следующем:

1. Матрица сети разбивается на блоки.
2. Для всех допустимых типов блоков в каждом блоке вычисляется ошибка (несоответствие).
3. Для каждого блока выбирается тип блока с минимальной ошибкой.
4. Вычисляется общая ошибка матрицы (сети) как сумма ошибок всех блоков.

Существует несколько видов обобщенного блокмоделинга: бинарный (binary), взвешенный (valued), имплицитный (implicit) блокмоделинг и блокмоделинг однородности (последние три могут быть

применены и к взвешенным сетям).

Процедура обобщенного блокмоделинга различна для бинарных и взвешенных сетевых данных. В бинарных сетях значения связей не учитываются, а рассматривается только наличие или отсутствие связей. Ошибка вычисляется как количество связей, не совпадающих со связями идеальных блоков. Способ определения этих связей зависит от типа блока. Наиболее распространенными идеальными блоками являются NULL, COMPLETE, REGULAR, ROW-REGULAR и COLUMN-REGULAR. Для взвешенных сетей значения линий рассматриваются с заданным пороговым значением  $m$ , при превышении которого связи считаются значимыми. Несоответствия измеряются как отклонения от 0 (связь не должна присутствовать) или  $m$  (связь должна присутствовать). Отклонение от  $m$  вычисляется только в том случае, если значение связи меньше  $m$ . Основная проблема заключается в определении наиболее подходящего значения  $m$ . В этом может помочь рассмотрение распределений либо значений ячеек, либо значений функции  $f$  по строкам или столбцам. Бинарный блокмоделинг является частным случаем взвешенного блокмоделинга.

Стохастический блокмоделинг основан на некоторой вероятностной модели. Предполагается наличие базовой статистической модели и ее оценка путем максимизации некоторой меры правдоподобия. Модель позволяет делать статистические выводы. Стохастические блокмодели (SBM) являются все более популярным классом моделей в статистическом анализе графов или сетей. Они могут использоваться для выявления или понимания (латентной) структуры сети, а также для кластеризации. Такая модель строится путем одновременного взятия каждой пары узлов и моделирования (неориентированного) ребра между ними. Вероятность наличия или отсутствия такого ребра не зависит от вероятности любой другой пары узлов.

**3.10.3.2 2.2. Детерминированный и стохастический блокмоделинг: подходы к анализу динамических сетей** В обзоре подходов к детерминированному и стохастическому динамическому блокмоделингу сетей [90] рассматривались недавно предложенные подходы к блокмоделингу. Критерии отбора исключали подходы, которые еще не реализованы в R или MATLAB, а также подходы, реализованные таким образом, что во всех временных точках предполагается один и тот же тип блокмодели. В список были включены и те подходы к блокмоделингу, которые не были разработаны специально для блокмоделинга динамических сетей, но могут быть использованы для этой цели. Итоговый список рассмотренных подходов представлен в Табл. 1 (всего 11 подходов). Эти подходы характеризуются двумя признаками: типом характеристики подхода к блокмоделингу (стохастический или детерминистический) и способом обработки зависимостей между временными точками (модели пространства состояний, зависимые кластеры, связанные сети, отсутствие зависимостей). Для всех видов моделирования временных зависимостей разработаны только версии стохастического подхода к блокмоделингу, а детерминистические подходы ( $k$ -means и обобщенный блокмоделинг) учитывают временную зависимость только через связанные сети.

Ниже эти подходы рассмотрены в зависимости от типа блокмоделинга и моделирования временных зависимостей, которые они предполагают.

Name	Abbreviation	Reference	Time				Is considered?
			Blockmodeling approach	dependency	Further modelling		
Dynamic stochastic blockmodels for time-evolving social networks (with state-space models)	DSBMwSSM	(Xu and Hero, 2014)	Stochastic	State-space models			YES
Stochastic block transition models for dynamic networks (with state-space models)	SBTMwSSM	(Xu, 2015)	Stochastic	State-space models			YES
Statistical clustering of temporal networks through a dynamic stochastic block model	DSBM	(Matias and Miele, 2017)	Stochastic	Dependent clusters	Within-cluster tie probabilities are constrained to be equal across timepoints to control the label-switching problem	YES	

Name	Abbreviation	Reference	Time				Is considered?
			Blockmodeling approach	dependency modelling	Further restrictions		
An exact algorithm for time-dependent variational inference for the dynamic stochastic block model	EDSBM	(Bartolucci and Pandolfi, 2020)	Stochastic	Dependent clusters	The blockmodel/connectivity parameters cannot change in time	NO – do not allow the connectivity parameters to change in time	
Stochastic blockmodeling for multilevel networks	SBMfMLN	(Chabert-Liddell et al., 2021, Chabert-Liddell, 2022)	Stochastic	Dependent clusters	The approach was originally designed for two levels and therefore only two time points; still, extensions have been proposed to multiple levels/timepoints and only recently implemented	NO – not implemented for more than two time points at the time of preparing the simulations	
Block models for generalised multipartite networks	SBMfMPN	(Bar-Hen et al., 2020)	Stochastic	Linked networks			YES
Stochastic blockmodeling for linked networks	SBMfLN	(Škulj and Žiberna, 2022)	Stochastic	Linked networks			YES

Name	Abbreviation	Reference	Time			Is considered?
			Blockmodeling approach	dependency modelling	Further restrictions	
Generalised blockmodeling for linked networks	GBMfLN	(Žiberna, 2019)	Deterministic - Generalised	Linked networks		NO – too slow
K-means-based algorithm for blockmodeling linked networks	KMfLN	(Žiberna, 2020)	Deterministic - k-means	Linked networks		YES
Stochastic blockmodeling	SBM	OverviewFunke and Becker, 2019; implementation used Škulj and Žiberna, 2022)	Stochastic	none		YES
K-means based block-modeling	KM	(Žiberna, 2020)	Deterministic - k-means	none		YES

**3.10.3.3 Динамические стохастические блокмодели с использованием моделей пространства состояний (state-space models)** Отнесение моделей к *моделям пространства состояний* подразумевает, что значения параметров (плотности блоков) в текущей временной точке оцениваются “условно” по параметрам предыдущей временной точки.

Модель *Динамических стохастических блокмоделей* для развивающихся во времени социальных сетей (с моделями в пространстве состояний), или *Dynamic stochastic blockmodels for time-evolving social networks (with state-space models) (DSBMwSSM)*, представленная Xu и Hero (2014) [424], основана на статической стохастической блокмодели - классической стохастической блокмодели для одномодальных сетей, которую они используют в качестве статической модели для сети отдельных временных точек, а для моделирования динамики сети применяют модель пространства состояний. Модель разработана только для бинарных сетей. В основе модели лежит предположение о том, что блокмодель, точнее матрица внутриблочных вероятностей связей, эволюционирует как некое случайное блуждание. Поскольку вероятности связей ограничены в диапазоне от 0 до 1, для их преобразования в неограниченные вещественные числа используется логит-преобразование. Эта параметризация, в свою очередь, используется в режиме пространства состояний: при переходе от времени  $t - 1$  к времени  $t$  эти параметры сначала преобразуются на основе модели перехода состояний, а затем к преобразованным состояниям

добавляется шум многомерного нормального процесса с нулевым средним значением. При этом используется интерактивная оценка, то есть при блокмоделировании временной точки  $t$  рассматриваются только временные точки от 1 до  $t$ .

Одним из недостатков подхода является то, что матрица преобразования состояний и ковариационная матрица шума процесса обычно неизвестны. Последняя должна задаваться пользователем и является особенно проблематичной, так как существенно влияет на результаты. Фактически для определения этой матрицы необходимо установить два параметра - дисперсии (при прочих равных условиях) и ковариации. Одним из вариантов является сеточный поиск их разумных комбинаций, однако это увеличивает временную сложность, а также неясно, какой критерий выбора модели лучше использовать для получения результатов. Лог-правдоподобие (log-likelihood) всей модели не существует (из-за использования модели пространства состояний для моделирования динамики) и, следовательно, не может быть использовано для этой цели. При оценке разбиения для времени  $t$  фактически максимизируется не логарифмическое правдоподобие, а апостериорная плотность состояния. Как отмечают авторы, “процедура вывода состоит из расширенного фильтра Калмана (EKF) [3], дополненного стратегией локального поиска”. Процедура вывода является интерактивной, то есть при измерении оценки состояния в момент времени  $t$  используется только информация до момента времени  $t$  (включительно).

Позже Xu (2015) [423] расширил эту модель, назвав ее Стохастической моделью блочного перехода для динамических сетей, или Stochastic block transition model for dynamic networks (SBTMwSS), в которой вероятность присутствия связи зависит уже не только от блока, в котором эта связь находится, но и от того, присутствовала ли она в предыдущий момент времени. То есть если в другой модели для бинарных сетей существует одна вероятность, определяющая вероятность наличия связи в блоке, то в данной модели их две, одна из которых используется, если связь не присутствовала в предыдущих временных точках, а другая - если присутствовала. Естественно, что для первой временной точки используется только одна модель. В остальном модель очень похожа на DSBMwSSM, предложенную в работе Xu и Hero (2014) [424], включая процедуру оценки.

**3.10.3.4 Динамические стохастические блок-модели с управлением переключения лейблов (control for label switching)** Следующий подход к динамическим стохастическим блок-моделям основан на условных кластерных вероятностях (*conditional cluster probabilities*): кластерные вероятности в текущей временной точке зависят от принадлежности к кластеру в предыдущей временной точке (точках); модели предполагают контроль переключения лейблов.

Этот подход был разработан Матиасом и Миле [260], которые предложили динамическую стохастическую блокмодель, позволяющую контролировать переключение лейблов, т.е. они гарантируют, что один и тот же кластер имеет один и тот же лейбл во всех временных точках. Их подход к Статистической кластеризации временных сетей с помощью динамической стохастической блокмодели, или Statistical clustering of temporal networks through a dynamic stochastic block model (DSBM), основан на Стохастической блокмодели, или Stochastic blockmodel (SBM); точнее, каждая временная точка моделируется как SBM. Вероятности связей (для бинарных сетей или, в более общем случае, параметры, определяющие распределение вероятностей значений связей) в заданный момент времени зависят только от латентных групп/кластеров, то есть только от разбиения в этот момент времени.

В принципе, это означает, что параметры связности (блокмодель) могут быть различными в разных временных точках, хотя изменение параметров связности во времени ограничено. Временные точки связаны между собой тем, что вероятности кластеров для каждой временной точки (кроме первой) зависят от принадлежности к кластеру в предыдущей временной точке. Как отмечают Матиас и Миле, один из подходов обусловлен необходимостью контроля проблем переключения лейблов в разных временных точках, то есть они хотят обеспечить, чтобы один и тот же кластер имел один и тот же “лейбл” во всех временных точках. Они решают проблему переключения лейблов путем поиска групп, которые имеют стабильное поведение внутригрупповой связности. Это достигается путем ограничения параметров связности диагональных блоков (которые моделируют связи внутри групп) на постоянство в разных временных точках. Вероятности внутригрупповых связей фиксируются во времени. Как отмечают авторы, такой контроль за переключением лейблов является ключевым отличием от модели DSBMwSSM, предложенной в работе Xu и Hero (2014).

Для оценки параметров используется приближенная процедура [36], основанная на алгоритме вариационного ожидания-максимизации, или variational expectation–maximisation algorithm (VEM). В работе Bartolucci and Pandolfi (2020) [36] также предложен точный алгоритм, оптимизирующий ту же целевую функцию - *Точный алгоритм зависящего от времени вариационного вывода для динамической стохастической блочной модели, или Exact algorithm for time-dependent variational inference for the dynamic stochastic block model (EDSBM)*, однако их подход ниже не рассматривается, поскольку они не допускают изменения параметров связности во времени, что является частью критерия включения при выборе подходов.

Моделью, аналогичной модели Matias и Miele (2017) [260], является модель, предложенная Chabert-Liddell et al. (2021) [64], *-Стохастический блокмоделинг для многоуровневых сетей, или Stochastic blockmodeling for multilevel networks (SBMfMLN)*. Эта модель явно предназначена не для динамических сетей, а для многоуровневых сетей. Модель очень похожа на модель Matias и Miele (2017) с двумя основными отличиями. Во-первых, в ней не решается проблема переключения лейблов, что не актуально для многоуровневых сетей. Изначально она разрабатывалась только для двухуровневых сетей (и поэтому могла использоваться только для темпоральных сетей с двумя временными точками). Позднее [65] модель была обобщена на любое число уровней и, следовательно, на динамические сети, измеряемые в любое число временных точек. Однако это обобщение еще не было реализовано. Как и в модели Matias и Miele (2017), уровни сети (в данном случае) связаны между собой путем установления зависимости вероятностей кластеров второго уровня от принадлежности к соответствующей единице первого уровня.

**3.10.3.5 Стохастический блокмоделинг для многосторонних или связанных сетей** Подходы, представленные в данной группе, предполагают наличие связанных и многосторонних сетей: совокупность как минимум двух одномодальных сетей и одной двумодальной сети, связывающей эти одномодальные сети. В контексте динамических сетей двумодальные сети “связывают” одни и те же единицы из разных временных точек. Такая сеть блок-моделируется как единая сеть (с ограничением, что узлы из разных одномодальных сетей не могут смешиваться).

Представленные здесь подходы разработаны не специально для динамических сетей, а для того, что Бар-Хен и др. [35] называют “обобщенными многосторонними сетями” (“generalised multipartite networks”), а Жиберна [446] - “связанными сетями” (“linked networks”). Этот тип сетей можно рассматривать

как совокупность одномодальных и двумодальных сетей, где сети имеют одинаковые наборы единиц. В качестве альтернативы (более подходящей для динамического анализа сетей) их можно рассматривать как совокупность одномодальных сетей, связанных с двумодальными сетями. Эта структура данных может быть использована для представления одномодальных сетей, измеренных в нескольких временных точках. В этом случае одномодальные сети, измеренные в нескольких временных точках, очевидно, представляют собой одномодальные сети (по одной для каждой временной точки), а двумодальные сети используются для “соединения” одних и тех же единиц в разных (обычно последовательных) временных точках (как показано на fig. 7).

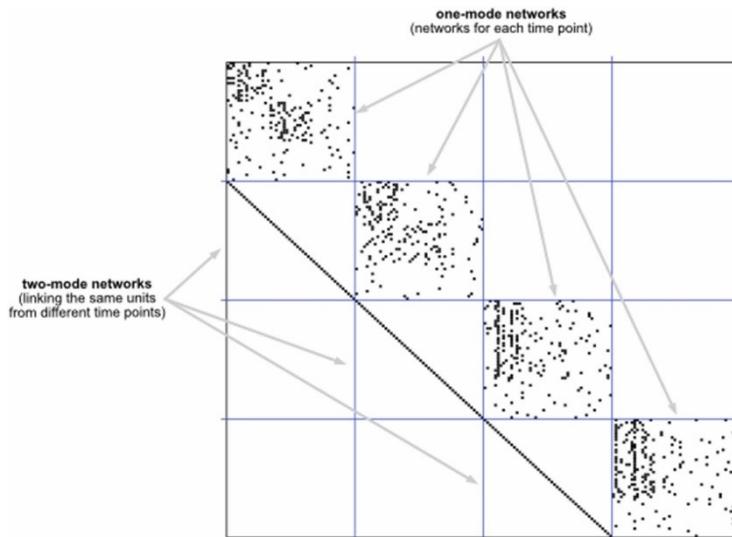


Figure 7: Иллюстрация одномодальных и двумодальных сетей, представляющих собой одномодальные сети, измеренные в разные моменты времени

Другой взгляд на связанную сеть состоит в том, что это просто гигантская сеть с различными наборами единиц, где требуется найти такие разбиения всех наборов единиц, чтобы кластеры состояли только из единиц одного набора, при этом при разбиении данного набора учитывается вся сеть, а также разбиения других наборов. В случае динамических сетей каждое множество будет представлять единицы, присутствующие в одной временной точке. Связи внутри множеств представляют собой отношения между наблюдаемыми единицами в данной временной точке, а связи между множествами представляют собой “однотипные” (“same-unit”) отношения (т.е. связывают одну и ту же единицу в соседних временных точках).

Преимуществом данного подхода является его гибкость, поскольку он подходит не только для динамических сетей, но и для других типов связанных сетей, например, многоуровневых или динамических многоуровневых сетей. Недостатком является то, что он не приспособлен специально для динамических сетей. Проблема заключается в том, что двумодальные “однотипные” сети имеют некоторые ограничения, которые не учитываются стохастической блокмоделью, а именно: каждый блок может иметь не более одной связи. Кроме того, эти подходы не решают проблему переключения лейблов, что также можно рассматривать как преимущество, поскольку это означает, что они не накладывают никаких ограничений на блокмодель или ее изменения во времени и не требуют одинакового количества кластеров во всех временных точках.

Два подхода, которые мы рассматриваем в этом разделе, - *Блокмодели для обобщенных*

*многосторонних сетей, или Block models for generalised multipartite networks (SBMfMPN)* Бар-Хена и др. [35] и *Стохастический блокмоделинг для связанных сетей, или Stochastic blockmodeling for linked networks (SBMfLN)* Шкуля и Жиберна [357] - очень похожи и оценивают практически одну и ту же модель, с тем лишь исключением, что в последней модели также предусмотрена возможность взвешивания. Кроме того, они используют разные стратегии оценивания. Подход SBMfMPN, предложенный Bar-Hen et al. (2020) [35], оценивает модель с помощью алгоритма максимизации вариационного ожидания (VEM), который инициализируется либо на основе разбиения, полученного с помощью иерархической кластеризации, либо с помощью отдельного блокмоделинга соответствующих подсетей (отметим, что важной частью процедуры оптимизации является также исследование разбиений, полученных путем разделения или слияния выбранных кластеров на предыдущей итерации). В подходе SBMfLN, предложенном Škulj и Žiberna (2022) [357], используется алгоритм максимизации классификационного ожидания (CEM), инициализируемый множеством случайных разбиений.

**3.10.3.6 Подходы детерминистического блокмоделинга для связанных сетей** До представленных выше подходов для связанных сетей были разработаны два нестохастических подхода к блокмоделингу: *Обобщенный блокмоделинг, или Generalised blockmodeling* и *Блокмоделинг методом k-средних, k-means blockmodeling*.

Обобщенный блокмоделинг был впервые разработан для многоуровневых сетей [445], но было показано, что этот же подход может быть использован и для других типов связанных сетей, в том числе динамических - *Обобщенный блокмоделинг для связанных сетей, или Generalised blockmodeling for linked networks (GBMfLN)* [446]. Обобщенный блокмоделинг — это очень гибкий подход для блокмоделинга бинарных и взвешенных сетей. Его гибкость обусловлена тем, что он может находить разделения на основе различных типов эквивалентности, включая структурную, регулярную и обобщенную эквивалентность [39]. Последний тип эквивалентности позволяет задавать различные типы блоков, где каждый тип блока представляет собой различное ожидание модели связей между двумя кластерами. В силу своей общности он использует для оптимизации алгоритм перемещения (relocation), что приводит к тому, что он становится слишком медленным для большинства связанных сетей, которые (по своей природе являясь комбинацией различных сетей) больше, чем типичные одномодальные сети.

Другим подходом является *Обобщенное блокмоделирование однородности суммы квадратов, Sum of squares homogeneity generalised blockmodeling* [443], которое минимизирует сумму внутриблочных квадратичных отклонений от среднего (для всех блоков сети) при использовании структурной эквивалентности. Этот же критерий минимизируется и в рамках *Блокмоделинга методом k-средних для связанных сетей, или k-means blockmodeling for linked networks (KMfLM)*. Поскольку подход k-means значительно быстрее обобщенного блочного моделирования и, следовательно, достаточно эффективен при работе со связанными сетями, рассматривался только этот нестохастический подход.

**3.10.3.7 Отдельный блокмоделинг для каждой временной точки** Выше были представлены подходы, которые могут быть использованы для блокмоделинга сетей во всех временных точках одновременно с учетом информации о соседних/других временных точках. Однако наиболее простым подходом к блокмоделингу snapshot-сетей является блокмоделинг сети в каждой временной точке отдельно (примеры см. в [89, 105, 106]. Жиберна [446] называет такой подход “раздельным анализом” и

утверждает, что практически он является необходимым первым (по крайней мере, исследовательским) шагом в блокмоделинге любой связанной, а значит, и динамической сети. Такие модели не предполагают моделирования зависимости от времени.

Для этого можно использовать любой подход к одномодальному блокмоделингу, но мы рассматриваем только частные случаи подходов к динамическим сетям. Поэтому для сравнения различных подходов мы используем только *Блокмоделинга методом k-средних для одномодальных сетей*, или *k-means blockmodeling for one-mode networks - KM* [447] (детерминистический подход) и *Стохастический блокмоделинг, Stochastic blockmodeling - SBM* ([20], [148], [303], [366]) (стохастический подход).

**3.10.3.8 Детерминированный и стохастический блокмоделинг: подходы для анализа ненаправленных сетей** В данной части рассмотрены только те подходы в области динамического блокмоделинга сетей, которые могут быть использованы для блокмоделинга ненаправленных динамических сетей. Случай ненаправленных сетей интересен для исследования тем, что модели могут быть применены к эмпирическим данным о сотрудничестве между учеными, основанном на их соавторстве.

Для сетей соавторства характерны некоторые особенности, такие как симметричные связи, наличие новичков и ушедших. Анализ сетей соавторства [89] показал, что в них присутствует значительная доля новичков и аутсайдеров, что приводит к образованию относительно неустойчивых кластеров в структуре глобальной сети. Была обнаружена типичная структура “мульти-ядро - полупериферия - периферия”, причем этот тип блокмодели остается относительно стабильным во времени. Этот фактор стабильности становится решающим при выборе между различными подходами к блокмоделингу. Для анализа динамики сетей соавторства словенских исследователей ставился вопрос о том, какой подход к блокмоделингу следует использовать. Этот вопрос может быть решен в ходе имитационного исследования.

Не все подходы, представленные в Табл. 1, могут быть использованы для блокмоделинга динамических ненаправленных данных. Были рассмотрены отдельные подходы к блокмоделингу, которые могут быть использованы при анализе динамических сетей, а также отдельные подходы к блокмоделингу (если сети рассматриваются только в одной временной точке), таким образом, что сети для каждой временной точки блок-моделируются отдельно:

1. Стохастический блокмоделинг - SBM (Mariadassou et al., 2010)
2. Блокмоделинг на основе К-средних - KM [447]
3. Байесовский стохастический блокмоделинг - BSBM [303].

Также применялись подходы, разработанные для анализа динамических сетей. Подходы *Dynamic stochastic blockmodels for time-evolving social networks (with state-space models) - DSBMwSSM* [424] и *Stochastic block transition models for dynamic networks (with state-space models) - SBTMwSSM* [423] не были включены в сравнение, поскольку они не позволяют учитывать входящие и выходящие узлы. Для анализа были выбраны следующие подходы:

4. Статистическая кластеризация временных сетей с помощью динамической стохастической блокмодели - DSBM [260]

5. Стохастический подход блокмоделинга для анализа многоуровневых сетей - SBMfML [65]
6. Байесовский стохастический блокмоделинг - BSBMfLN [303]
7. Блокмодели для обобщенных многосторонних сетей - SBMfMPN [35]
8. Стохастический блокмоделинг для связанных сетей - SBMfLN [357]
9. Алгоритм на основе К-средних для блокмоделирования связанных сетей - KMfLN [447]

*Стохастический подход блокмоделинга для анализа многоуровневых сетей, или Stochastic blockmodel approach for the analysis of multilevel networks - SBMfML [65]* очень похож на *Динамический стохастический блокмоделинг, или Dynamic Stochastic Blockmodeling*, предложенный Matias & Miele (2017) [260]. Количество групп не фиксировано для всех временных точек, а диагональные блоки могут меняться со временем в подходе *Стохастического блокмоделинга для многоуровневых сетей, Stochastic blockmodeling for multilevel networks - SBMfMLN*. В подходе *Динамического стохастического блокмоделинга*, предложенного Matias & Miele (2017), предполагается, что плотность диагональных блоков не должна меняться со временем, однако в подходе Chabert-Liddell (2022) [65] это не так. Кроме того, Matias & Miele также предполагают, что количество кластеров одинаково во всех временных точках. В подходе *Байесовского стохастического блокмоделинга, Bayesian stochastic blockmodeling - BSBM* [303] предполагается распределение связей Пуассона, что хорошо работает, особенно для разреженных сетей.

### **3.10.4 Сравнение подходов к динамическому блокмоделингу: результаты имитационных исследований**

**3.10.4.1 Сравнение детерминированных и стохастических подходов к динамическому блокмоделингу** Для сравнения подходов для направленных сетей было проведено имитационное исследование методом Монте-Карло, которое было представлено в работе Цугмаса и Жиберна [90]. Целью исследования было определение различий в эффективности различных подходов к блокмоделингу динамических сетей, определяемых как сети моментальных снимков, наблюдаемых в нескольких временных точках или периодах времени на одном и том же наборе единиц и с одним и тем же типом связи между единицами.

Выбранные подходы к блокмоделингу оценивались с учетом нескольких факторов, т.е. характеристик сети. Рассмотренные подходы к блокмоделингу применялись к синтезированным сетям с различными свойствами, такими как размер (48 или 96 узлов), устойчивость групп (стабильность – членство в группе; четыре случая: обмен 0%, 16%, 33%, 100% узлов), разница между плотностями блоков (нулевой и полной) и переходами между типами блок-моделей (две сплоченные и одна периферийная группы, где могут измениться два или четыре внедиагональных блока, а также один или два внедиагональный и один или два диагональный блок). Связи между группами обычно не являются случайными и возникают под действием некоторых социальных механизмов ([2], [373]), таких как взаимность, популярность, транзитивность, а также механизм исходящих партнеров (outgoing-shared partner) - тенденция к созданию связей с теми, кто “нравится другим”. В эксперименте были возможны два сценария: 1) механизмы локальной сети не учитываются, то есть связи генерируются совершенно случайным образом, и 2) механизмы учитываются, и получаются различные типы блок-моделей. Во всех генерируемых сетях предполагалось, что известно истинное число групп – три группы. В исследовании предполагалось, что все узлы одинаковы во всех временных точках (т.е. нет входящих и выходящих

узлов). Для каждой комбинации факторов было сгенерировано десять сетей.

Алгоритм генерации сетей является итерационным, поэтому пришлось задать число итераций равным 2500. Подход к генерации сетей показан на fig. 8 (подробнее см. [90]).

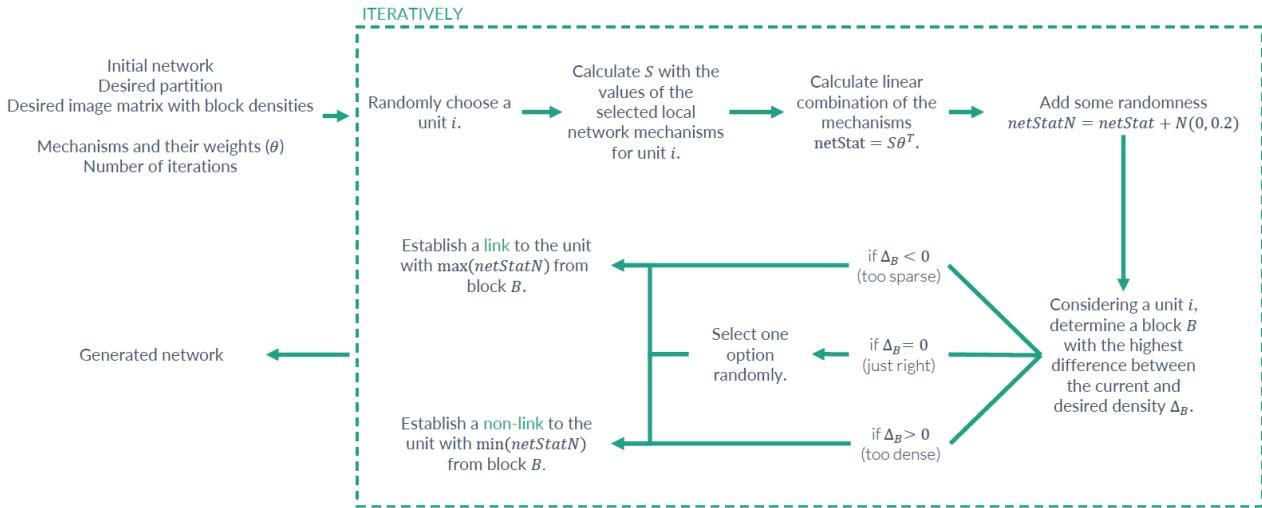


Figure 8: Итерационный подход к генерации сетей: динамический блокмоделинг сетей

Сначала проводился блокмоделинг сетей с использованием подходов для раздельного блокмоделинга сетей, наблюдавшихся в разные временные точки (раздельный блокмоделинг). Здесь использовались стохастический подход к блокмоделингу (SBM), оцененный с помощью алгоритма СЕМ [357], и подход к блокмоделингу на основе k-средних [447]. Также применялись подходы блокмоделинга для динамических, связанных и многоуровневых сетей. Подходы к блокмоделированию оценивались по скорректированному индексу Рэнда [318], который измеряет сходство истинных разделений и разделений, полученных с помощью блокмоделирования.

В результате анализа было показано, что локальные сетевые механизмы могут влиять на решения блокмоделинга. Например, в реализованном исследовании средние значения ARI были ниже при учете механизмов по сравнению с тем, когда связи внутри блоков генерировались случайным образом. Другим важным фактором является изменение типа блокмодели во времени. Более выраженные изменения типа блокмодели приводят к более низким средним значениям ARI.

Низкая устойчивость групп во времени приводит к худшим результатам подходов блокмоделинга для динамических сетей. Более низкие результаты наблюдаются в тех случаях, когда разбиение из первой временной точки и разбиение в последней временной точке являются независимыми. В этом случае блокмоделинг сетей, наблюдавшихся в разные моменты времени по отдельности, может дать приемлемые результаты. Однако при наличии хотя бы некоторой зависимости между разделениями рекомендуется рассмотреть подходы блокмоделинга для динамических сетей.

В целом ожидается, что статистическая кластеризация временных сетей с помощью динамической стохастической блокмодели - DSBM [260] даст наилучшие результаты, если тип блокмодели одинаков во всех временных точках (не изменяется во времени). Такой подход к моделированию блоков предлагается независимо от устойчивости разделений и возможных социальных механизмов, влияющих на связи внутри блоков.

Если предполагается, что тип блокмодели будет меняться во времени, то оптимальными

могут оказаться блокмодели для обобщенных многосторонних сетей - SBMfMPN [35] и динамические стохастические блокмодели для эволюционирующих во времени сетей - SBTMwSSM [423], причем в большинстве случаев предпочтение следует отдать SBMfMPN.

Поскольку характеристики сети (тип блокмодели, устойчивость группы, локальные механизмы сети) существенно влияют на эффективность подходов к блокмоделингу, настоятельно рекомендуется учитывать предварительные знания о сети и возможные предварительные (например, отдельные блок-модели) анализы. Кроме того, учет различных начальных разбиений может привести к лучшим значениям критериальной величины (а значит, и к более подходящему решению) и способствовать сходимости моделей. Рекомендуется использовать различные начальные разделения (например, из отдельных анализов) и сохранять решение с наилучшим значением критерия.

В дальнейших исследованиях было бы полезно оценить подходы к блокмоделированию в ситуациях с более неоднородным количеством кластеров в сетях, более неоднородными размерами кластеров в сетях, а также когда истинное количество кластеров неизвестно. На решение в рамках блокмоделинга также может повлиять количество рассматриваемых временных точек и/или время между временными точками. Поскольку это напрямую связано с динамикой сети, т.е. с устойчивостью типа блокмодели и разбиения, предполагается, что подходы к блокмоделингу динамических сетей могут принести меньше пользы по сравнению с раздельным анализом, если время между временными точками больше, исходя из того, что социальные сети имеют тенденцию изменяться во времени. Несмотря на то что в данном исследовании были получены сети, напоминающие характеристики реальных сетей, в будущем следует предпринять попытки оценить достоверность подходов блокмоделинга на реальных сетях.

**3.10.4.2 Сравнение детерминированных и стохастических подходов к динамическому блокмоделингу ненаправленных сетей** После проведенного исследования было рекомендовано на дальнейших этапах сравнения подходов динамического блокмоделинга учитывать возможность присоединения и выхода единиц из сети с течением времени, что часто происходит в динамических социальных сетях реального мира. Было решено провести аналогичную оценку для ненаправленных сетей, для чего алгоритм генерации сетей должен быть соответствующим образом адаптирован.

В имитационном исследовании, выполненном Цугмасом и Жиберна, в качестве случайных эффектов рассматривались те же сценарии размера сети (48 против 96) и стабильности (перехода группы) (0% против 33% против 66% против 100%), что и в предыдущих случаях, но с учетом изменений в механизмах, а также новичков и ушедших.

Исследовалось четыре сценария для новичков и ушедших (выбывших):

- Новые участники (0 против 50%, по сравнению с размером сети в первой временной точке)
- Выбывшие (0 против 20%, по сравнению с размером сети в первой временной точке)

Было обнаружено, что приход новых участников не изменял размер сети, а уход - уменьшал. Когда новички присоединяются к группе, у них нет существующих связей с другими участниками, и группа, к которой они присоединяются, определяется с вероятностью, пропорциональной ее размеру. Это реалистично для сетей соавторства или исследовательских групп, где большие группы имеют больше шансов привлечь новых коллег, возможно, из-за наличия финансовых ресурсов. При этом все группы в

сети имеют одинаковую вероятность выхода, за двумя исключениями. Чтобы избежать появления очень маленьких групп, ни одна единица из группы не могла покинуть ее, если группа состоит из пяти или менее единиц. Кроме того, новички также не могут покинуть сеть в течение одного и того же периода времени.

Что касается *механизмов*, то мы либо использовали механизмы, либо учитывали такие механизмы, как популярность, транзитивность и ассортативность (склонность к созданию связей с теми, кто имеет схожие показатели in-degree). В случае популярности есть предпочтение устанавливать связи с теми, кто очень популярен. В случае с ассортативностью есть предпочтение связываться с теми, кто имеет схожий уровень популярности. Также учитывается транзитивность - склонность узла устанавливать связи с теми, кто нравится связанным с ним узлам.

В качестве *фиксированных эффектов* использовалось одинаковое количество групп (3 группы), плотность блоков (5% в нулевых блоках и 25% в полных блоках). Что касается типов блокмоделей, то они также различны, но количество возможных типов блок-моделей ограничено из-за наличия симметричных связей. Начальный тип блокмодели аналогичен тому, который был использован при анализе динамического блокмоделинга сети, с двумя сплоченными группами и одной периферийной группой. Далее мы исследуем три сценария окончательной структуры глобальной сети. В первом случае ничего не меняется, то есть структура глобальной сети остается прежней. Во втором сценарии происходят два слияния, в результате чего появляются связи между первой и второй группой и третьей группой. В третьем сценарии изменяется один диагональный и один внедиагональный блок, при этом исчезают связи внутри блоков второй группы и появляются связи между первой и второй группами.

*Алгоритм генерации сетей* аналогичен описанному выше (подробнее см. [90]), с дополнительным шагом задания количества пришедших и ушедших, а также минимального размера кластера. Процедура генерации сетей, в которой было использовано 50 000 итераций, показана на fig. 9.

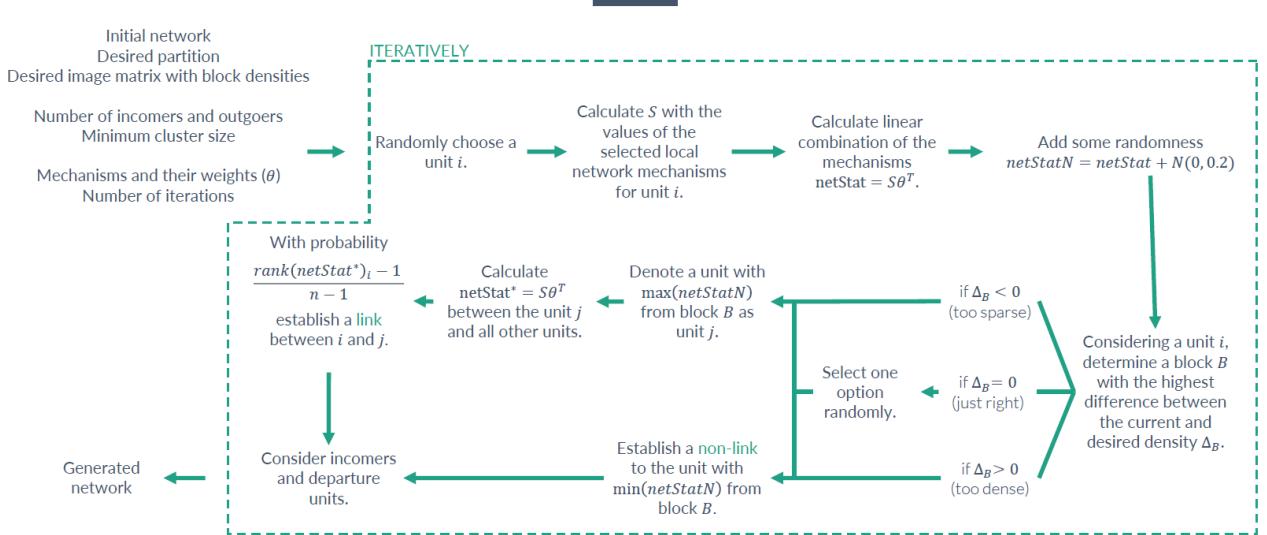


Figure 9: Итерационный подход для генерации сетей: динамический блокмоделинг ненаправленных сетей

Алгоритмы запускались дважды: с начальным разбиением по умолчанию, а затем с начальным разбиением, полученным в результате раздельного блокмоделинга. Для динамических сетей для каждого подхода получено два набора результатов, причем для каждого подхода, примененного к динамическим сетям, был оставлен и далее проанализирован тот, который имеет наилучшее значение оптимизированной

критериальной функции. Для сравнения истинного и полученного с помощью этих подходов разбиения для динамических сетей в каждой временной точке применен скорректированный индекс Рэнда (ARI) (для измерения степени сходства расчетных разбиений с истинными). Были вычислены среднее значение ARI по всем временным точкам и далее проинтерпретированы полученные результаты.

На основании полученных результатов можно дать следующие рекомендации:

1. Стохастический подход блокмоделинга для анализа многоуровневых сетей - SBMfML [65] является очень надежным выбором, когда нет входящих и выходящих узлов.
2. Стохастический блокмоделинг для многосторонних сетей - SBMfMPN [35] - безопасный выбор, когда есть входящие и выходящие узлы. Однако следует иметь в виду, что иногда могут возникать проблемы со сходимостью, хотя это случается редко.
3. Алгоритм блокмоделинга связанных сетей на основе K-средних - KMfLN [447] в целом эффективен, но может быть превзойден другими подходами. Лучше всего он работает, если связи внутри блоков устанавливаются случайным образом. Его можно рассматривать как надежный вариант, хотя могут существовать альтернативы, дающие лучшие результаты.
4. Динамическая стохастическая блокмодель - DSBM [260] работает очень хорошо, если можно предположить, что связи в блоках или кластерах устанавливаются случайно, особенно в больших сетях и когда тип блокмодели не меняется. Однако при изменении типа блокмодели и больших размерах сети его производительность оказывается относительно ниже по сравнению с другими подходами. В таких случаях более эффективными могут оказаться другие подходы.

В дальнейшем эти подходы к блокмоделингу следует применять и оценивать на реальных сетях соавторства, учитывая значительный объем знаний о сетях сотрудничества в случае словенских исследователей.

В качестве общих рекомендаций следует отметить, что при начале анализа сети всегда целесообразно использовать предыдущие знания и проводить отдельные анализы. Целесообразно начинать с отдельных предварительных анализов для подтверждения знаний о сети. На эффективность подходов блокмоделинга могут влиять различные факторы. Вторая рекомендация - попробовать использовать различные начальные разбиения, например, начальное разбиение по умолчанию или из отдельных анализов. Следует оставить решение с наилучшим значением критерия. Кроме того, можно рассмотреть возможность анализа каждой временной точки отдельно и использовать эти результаты в качестве начальных разделов для динамического блокмоделинга. В некоторых случаях это может повысить эффективность подхода блокмоделирования.

### **3.10.5 Блокмоделинг для изучения сетей соавторства**

При анализе сетей соавторства многие ученые изучали преимущества сплоченных социальных структур, организованных в четко определенные, тесно связанные сообщества связанных между собой индивидов. Было показано, что блокмоделинг может быть эффективно использован для решения соответствующих исследовательских задач, разрабатываемых в области научного сотрудничества исследователей. Существует несколько типичных вариантов его использования, таких как:

1. Поиск кластеров единиц, имеющих эквивалентные (в некотором смысле) связи с другими единицами. Блокмоделинг может применяться для выявления исследователей, имеющих схожие

модели научного сотрудничества, и для выявления глобальной (сетевой) структуры научного сотрудничества, не накладывая при этом никаких предположений на глобальную сетевую структуру.

2. Поиск кластеров единиц в соответствии с некоторой заранее заданной структурой сети. Мы знаем, что структура глобальной сети имеет вид, например, ядро-периферия. Блокмоделинг может применяться для того, чтобы определить, какие исследователи относятся к ядру, а какие - к периферии.
3. Оценка соответствия заданного разбиения ко всей сети (и заданной заранее структуре). Можно предположить, что структура глобальной сети имеет вид “ядро-периферия”. Блокмоделинг применяется для оценки соответствия эмпирической сети структуре “ядро-периферия”. Также может существовать разбиение, определяемое принадлежностью к научному полю, и можно оценить, насколько это разбиение соответствует сплоченной сетевой структуре.

Вопрос о сплоченных социальных структурах может быть частично распространен на уровень долгосрочных и краткосрочных научных коллабораций. Однако в литературе еще недостаточно внимания уделяется факторам, влияющим на устойчивость коллаборативных связей, что включает вопросы динамики и стабильности полученных структур. Ниже рассмотрено несколько примеров применения блокмоделинга для изучения устойчивости и динамики сетей соавторства на примере словенской науки.

Кронеггер и др. [105] исследовали структуру сетей соавторства четырех научных дисциплин (физики, математики, биотехнологии и социологии) за четыре 5-летних периода (1986-2005 гг.). Один из ключевых выводов состоит в том, что независимо от научной дисциплины структура соавторства очень быстро консолидируется в мульти-ядро - полуперифирию - периферию. Термин “ядро” обозначает группу исследователей, которые систематически публикуются вместе друг с другом. Полупериферию состоит из исследователей, которые публикуются в соавторстве менее систематически, но имеют хотя бы одну публикацию в соавторстве с исследователями из данной дисциплины; периферия включает авторов, которые публикуются только как единственные авторы или с авторами, не входящими в границы определенной дисциплинарной сети [255]. В работе Kronegger et al. (2011) [105] также рассматривался вопрос об устойчивости ядер (групп исследователей) в контексте изучения эволюции блок-модельной структуры во времени с помощью визуального представления.

В более позднем исследовании Cugmas et al. (2016) [89] рассмотрена структура устойчивости сетей соавторства во времени. Процедура блокмоделинга была применена к дисциплинарным сетям соавторства словенских исследователей по 72 научным дисциплинам в двух временных периодах: 1991-2000 гг. и 2001-2010 гг. Изучалось, как меняется структура глобальной сети и сохраняется ли стабильность научного сотрудничества с течением времени. Целью исследования был анализ различий в стабильности и размере групп исследователей, работающих в соавторстве друг с другом (основных исследовательских групп), сформированных в дисциплинах, относящихся к естественным и техническим наукам, с одной стороны, и к социальным и гуманитарным наукам, с другой. В этом исследовании временная зависимость не учитывалась, и сети анализировались в двух временных точках.

Ядра были получены с помощью заранее специфицированной процедуры блокмоделинга, предполагающей структуру мульти-ядро - полупериферию - периферию. Устойчивость полученных ядер

измерялась с помощью модифицированного скорректированного индекса Рэнда.

Предполагаемая структура мульти-ядро - полупериферия - периферия подтвердилась во всех анализируемых дисциплинах. Средний размер полученных ядер выше во втором временном периоде, причем средний размер ядра больше в естественных и технических науках, чем в социальных и гуманитарных. Различий в средней стабильности ядра между естественными и техническими науками и социальными и гуманитарными науками не наблюдается. Однако если стабильность ядер определяется по разделению ядер, а не по доле исследователей, покинувших ядра, то при контроле характеристик сетей и дисциплин средняя стабильность ядер выше в дисциплинах из научных областей “Инженерные науки и технологии” и “Медицинские науки”, чем в дисциплинах гуманитарных наук.

В 2020 году были получены новые данные и результаты, продолжающие исследование еще на десять лет. Это позволяет проанализировать три временные точки, каждая из которых охватывает десятилетний период. Вопросы исследования касаются эффектов применения процедуры блокмоделинга с учетом временной зависимости и устойчивости полученной структуры. Поэтому в рамках имитационного исследования методом Монте-Карло было проведено сравнение подходов к анализу динамического блокмоделинга, в соответствии с которым были выбраны наиболее эффективные и безопасные для использования подходы.

В будущем эти подходы к блокмоделингу должны быть применены и оценены на реальных сетях соавторства, учитывая значительные знания о сетях сотрудничества в случае словенских исследователей.

### 3.10.6 Выводы

В данном разделе проводится систематизация различных подходов в области блокмоделинга и их сравнительный анализ для оценки эффективности применения при анализе динамических сетей. В качестве примера взяты сети соавторства, изменяющиеся во времени.

После рассмотрения общих особенностей применения методологии блокмоделинга и проблематики изучения динамических сетей рассматриваются особенности использования методологии блокмоделинга для анализа динамических сетей. Поскольку одной из основных особенностей анализа динамических сетей является проверка их устойчивости, рассматриваются также основные сведения о подходах к оценке устойчивости структуры блокмоделей на основе Рэнд-индекса. В работе рассматривается два подхода к блокмоделингу – детерминированный и стохастический, - и приводится разбор разработанных на основе этих подходов моделей, в т.ч. тех, которые применяются для анализа динамических сетей. На основе работы Цугмаса и Жиберны, в обзоре приведены основные результаты сравнения моделей с помощью имитационного моделирования методом Монте-Карло (метод симуляции), в т.ч. модели для исследования ненаправленных сетей в динамике. В завершение анализируется использование блокмоделинга для изучения ненаправленных сетей соавторства во времени в различных научных дисциплинах.

Подводя итог, можно сказать, что лучшими подходами к блокмоделингу, исходя из полученных результатов, являются:

- Динамический стохастический блокмоделинг - DSBM [260]
- Стохастический блокмоделинг для анализа многоуровневых сетей - SBMfML [65]
- Стохастический блокмоделинг для многосторонних сетей - SBMfMPN [35]

Динамический стохастический блокмоделинг используется для стабильных сетевых структур. В противном случае используется стохастический блокмоделинг для многоуровневых сетей, когда нет приходящих и уходящих узлов. При наличии входящих и уходящих узлов нужно рассматривать стохастический блокмоделинг для многосторонних сетей.

Эти рекомендации призваны упростить процесс принятия решений. Хотя в конкретных случаях они не всегда дают абсолютно оптимальные результаты, эти подходы минимизируют риск получения ошибочных или недостоверных результатов. При анализе реальных сетей часто бывает сложно определить, полностью ли выполняются все исходные предположения различных подходов. Поэтому выбор в пользу более безопасных подходов обеспечивает определенный уровень уверенности в результатах анализа.

### **3.11 Современные подходы в области статистического сетевого анализа и моделирования: модели SIENA, ERGM, tERGM**

#### **3.11.1 Введение**

Существует растущий спрос на реалистичные и интерпретируемые статистические модели для анализа сетей, и в частности для тех сетей, которые представлены в динамике. В контексте зависимых данных (тех, что нельзя назвать независимыми и случайно распределенными) были разработаны несколько подходов для статистического вывода (statistical inference) – к ним относятся иерархическое моделирование [227], временные ряды [54], пространственный анализ [411] и моделирование многомерных распределений с использованием копула-функций [138]. Однако ни один из этих методологических подходов не является достаточным для того, чтобы отразить сложную структуру и широкий диапазон зависимостей, которые мы наблюдаем в сетях.

Так, например, в области научного сотрудничества существует необходимость в разработке моделей, основанных на зависимых данных, для анализа сетей сотрудничества. Научное взаимодействие, социальная и когнитивная структура различных научных областей успешно изучаются в библиометрии и саентометрии (scientometrics) с помощью анализа временных библиометрических сетей – соавторства, цитирования, со-цитирования и библиометрических связей между авторами или группами авторов и библиометрическими сущностями, представленными в базе данных (работы, авторы, журналы, ключевые слова, организации, страны, и т.д.). В контексте анализа сетей сотрудничества модели позволяют нам понять, какие факторы стоят за образованием связи в сети, т.е. понять, что способствовало формированию данной структуры сотрудничества в академии.

С ранней работы Прайса [314] и работы Гарфильда [137], социологи представили несколько теорий, касающихся научного сотрудничества. Анализ саентометрии основан на эффекте Мэтью [314] и теории структуры малого мира [100], а также их применения к моделированию динамики сетей соавторства. Настоящий доклад будет сосредоточен на изучении потенциального применения экспоненциальных моделей случайных графов (ERGMs) и темпоральных экспоненциальных моделей случайных графов (TERGMs) в области библиометрического анализа.

### 3.11.2 Стохастический и детерминированный подходы к сетевому анализу

В области сетевого анализа стохастические методы представляют собой кульмиационные этапы аналитического процесса. Основой передовых аналитических процедур в сетевом анализе неизменно является детерминированный подход. Детерминированный подход к анализу сетей служит исходной базой, в рамках которой развиваются более сложные аналитические методики. В этом контексте мы классифицируем детерминированные подходы по трем основным направлениям: глобальные свойства, локальные свойства и разбиение на части (partitioning), каждое из которых позволяет по-разному взглянуть на структурные характеристики сети (см. fig. 10).

Глобальные свойства включают в себя фундаментальные атрибуты, позволяющие получить целостное представление о сети в целом. Эти свойства включают в себя различные аспекты, в том числе:

1. **Размер сети.** Этот параметр характеризует общее количество узлов или вершин в сети. Он дает фундаментальное представление о масштабе сети.
2. **Плотность.** Плотность определяет степень взаимосвязанности в сети. Она измеряет долю существующих связей по отношению ко всем возможным связям в сети.
3. **Централизация сети.** Централизация сети оценивает концентрацию центральных узлов в сети. Она позволяет определить, оказывают ли несколько узлов непропорционально большое влияние на взаимодействие в сети.
4. **Распределение степеней.** Показатель характеризует распределение степеней вершин в сети.
5. **Транзитивность** определяет склонность узлов к образованию кластеров или групп.
6. **Ассортативность и гомофиля.** Эти свойства изучают характер связей между узлами на основе общих характеристик или атрибутов. Ассортативность изучает склонность узлов со схожими характеристиками к соединению, а гомофиля – склонность узлов с общими характеристиками к взаимодействию.

Помимо описательной статистики, глобальные свойства позволяют получить ценные сведения о глобальной структуре сети, включая ее связность и наличие характерных конфигураций, таких как симметричные и асимметричные структуры “ядро-периферия”. Эти глобальные свойства тесно связаны с областью блок-моделирования.

Локальные свойства, напротив, позволяют проникнуть в микроструктуру сети и понять взаимодействие между отдельными узлами. Эти свойства включают в себя разнообразную описательную информацию о данной сети, в том числе:

1. **Меры центральности.** Центральные показатели, такие как степенная центральность (degree centrality), центральность близости (closeness centrality) и промежуточная центральность (betweenness centrality), отражают значимость отдельных узлов в сети и их роль в распространении информации или влияния. Изучение корреляций между различными центральностями позволяет выявить закономерности важности узлов.
2. **Коэффициент кластеризации.** Коэффициент кластеризации определяет склонность узлов к образованию кластеров или скоплений. Он дает представление о распространенности локальных структур сообщества в сети.

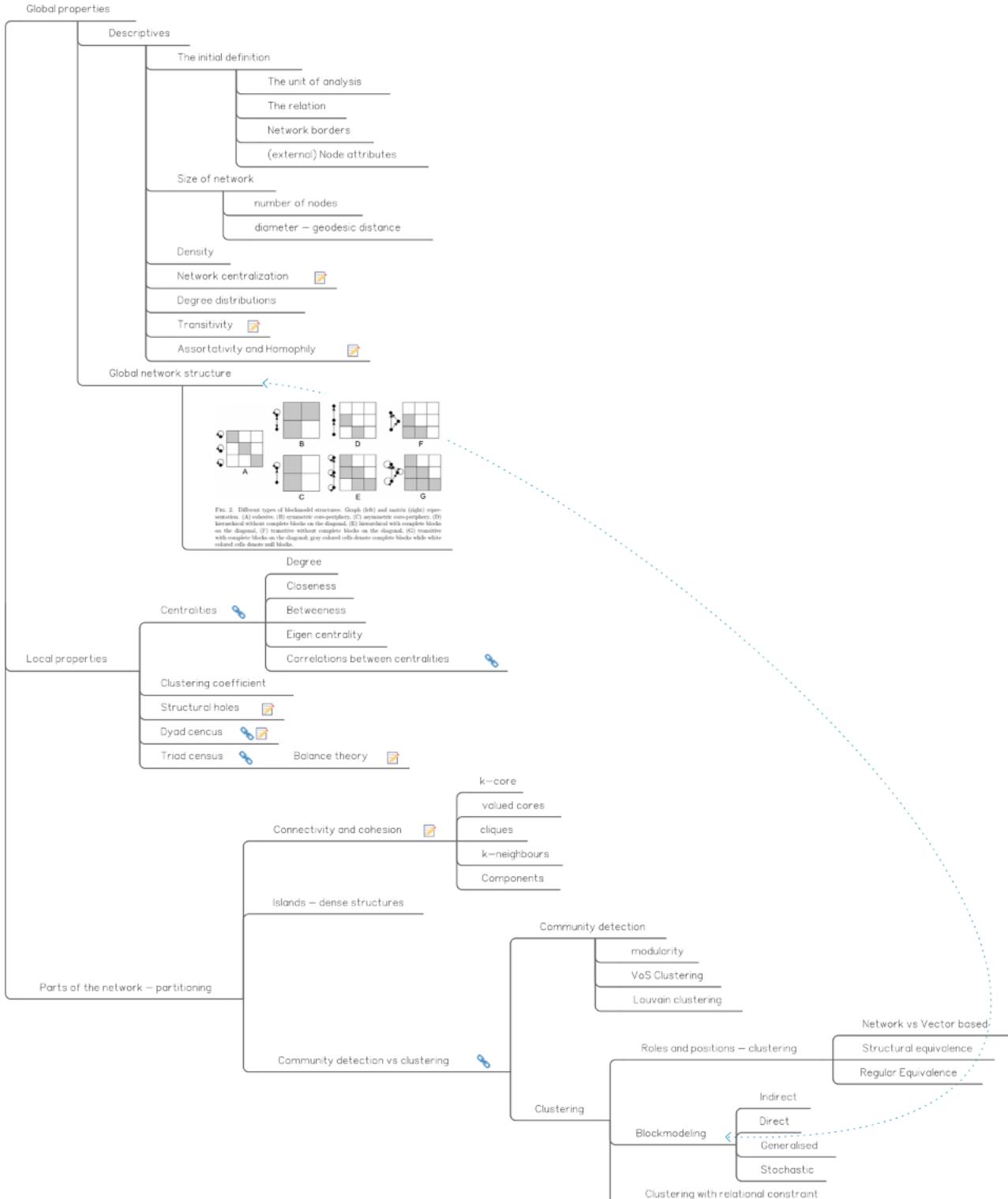


Figure 10: Детерминистские подходы к анализу сетей

3. *Структурные дыры.* Эта концепция изучает наличие брокерских возможностей в сети, когда отдельные лица или узлы служат мостами между разрозненными группами или кластерами.
4. *Dyad Census и Triad Census.* Переписи диад и триад предполагают категоризацию и подсчет конкретных сетевых подструктур, состоящих из двух или трех узлов соответственно. Эти метрики облегчают анализ структурных паттернов и мотивов в сети.
5. *Теория баланса.* Теория баланса изучает наличие сбалансированных или несбалансированных отношений в триадах узлов, внося свой вклад в понимание социальной динамики и стабильности сети.

Детерминированный подход также включает в себя разбиение сети (partitioning) – классификацию вершин сети таким образом, чтобы каждая вершина относилась ровно к одному классу или кластеру:

1. *Блоки связности и сплоченные подгруппы.* Социальные сети обычно содержат плотные скопления участников, которые взаимодействуют больше между собой, чем с другими участниками сети. Методы обнаружения сплоченных подгрупп включают k-ядра, ядра, клики, k-соседей и компоненты, которые выделяют в сети сплоченные группы. Общая гипотеза заключается в том, что люди, совпадающие по социальным характеристикам, будут взаимодействовать чаще, а люди, взаимодействующие регулярно, будут формировать общее отношение или идентичность [98].
2. *Острова.* Остров – это максимальная подсеть вершин, связанных между собой, значение которых больше, чем ребер, ведущих к вершинам вне такой подсети [98]. Другими словами, это плотно связанные друг с другом узлы, отражающие локально важные участки сети. Алгоритм для поиска островов доступен в программе Rajeck.
3. *Обнаружение сообществ и кластеризация вершин.* Различие между обнаружением сообществ (с использованием таких метрик, как модульность, VOS-кластеризация и кластеризация по методу Лувена) и кластеризацией (с использованием ролей, позиций, блок-моделирования и реляционных ограничений) предполагает различные точки зрения на выявление значимых подгрупп в сетях, каждая из которых подходит для решения конкретных аналитических задач [[98]].

Таким образом, детерминированный сетевой анализ представляет собой строгую основу для изучения сложных сетей. Его трехсторонняя структура включает в себя глобальные и локальные свойства, позволяющие понять структуру и динамику сети, а также разбиение сети на части, позволяющее выделить значимые подструктуры сети. Подход ориентирован на *статические* отношения между акторами. Он служит отправной точкой для анализа и закладывает основу для развития стохастических методов, предоставляя исследователям инструменты для всестороннего раскрытия многогранной природы сетевых систем.

Стохастический подход в своей основе опирается на результаты, полученные в рамках детерминированного подхода. В данной работе будут рассмотрены экспоненциальные модели случайных графов (ERGM), являющиеся одними из основных методов моделирования статических сетей. Модели ERGM служат универсальной аналитической основой, позволяющей исследовать различные сетевые явления (кластеризацию, гомофилю и другие структурные показатели, возникающие в результате сложного взаимодействия и поведения участников сети). Используя модели такого типа, мы можем выяснить, как и почему ученые сотрудничают друг с другом, и понять, что заставляет их работать вместе.

ERGM предлагает новый подход к моделированию состояния сети, отходя от традиционных методов регрессии. Вместо того чтобы предполагать независимость участников сети или связей, он рассматривает наблюдаемую сеть как один результат из многомерного распределения. Исследователи могут использовать ERGM для анализа сетей на основе гипотез, аналогичных тем, которые используются в классической регрессии (например, как ковариата влияет на результат), и при этом учитывать структуру или взаимозависимость сети в той мере, в какой они считают это целесообразным.

Однако, многие сети представлены в динамике, поэтому все большее признание получает необходимость выхода за рамки статических моделей и учета временной динамики. Во временном стохастическом подходе предполагается, что сетевые данные могут наблюдаться и измеряться в различные моменты времени, причем эти наблюдения не изолированы, а взаимосвязаны – они образуют последовательности, содержащие ценную информацию об эволюции сети. Возможные модели для динамических сетей представлены на Рисунке fig. 11.

		Theory	
Measurement	Cross-Sectional Data	Actor Based Models	Tie Based Models
	Panel Data	<b>SAOMs</b> [Snijders, 2005]	<b>(t)ERGMs</b> [Krivitsky and Handcock, 2014]
	Time-Stamped Data	<b>DyNAMs</b> [Stadtfeld and Block, 2017]	<b>REMs</b> [Butts, 2008]

Figure 11: Классификация основных стохастических подходов к анализу динамических сетей

**Акторно-ориентированные модели (SAOM).** Акторно-ориентированные модели – это класс временных стохастических моделей, в которых основное внимание уделяется отдельным участникам сети. SAOM, разработанные Снайдерсом [364], основаны на идее, что поведение и решения отдельных участников определяют изменения в сети с течением времени. Эти модели учитывают, как участники адаптируют свои связи в зависимости от своих характеристик и взаимодействия с другими участниками, что делает их ценными для понимания микроуровневой динамики развития сети. SAOM широко применяются в различных областях, включая социологию, организационное поведение и здравоохранение. SAOM представляют собой гибкую структуру для моделирования динамики социальных сетей и получения представления о механизмах, определяющих образование и распад связей с течением времени [367]. SAOM часто узнают по ее программной реализации, известной как SIENA.

**Модели, основанные на связях (TERGMs).** Временные модели экспоненциальных случайных графов (TERGM), предложенные П. Кривицким и М. Хандоком [217], используют другой подход, фокусируя внимание на связях. Эти модели рассматривают образование и распад связей с течением времени, исследуя глубинные механизмы, приводящие к изменениям в структуре сети. TERGM особенно полезны для отражения динамики и зависимостей на уровне связей.

**Диадические сетевые авторегрессионные модели (DyNAMs).** Диадические сетевые авторегрессионные модели, предложенные К. Штадтфельдом и его коллегами [374], сочетают в себе временное измерение и моделирование на основе акторов. В этих моделях изучается то, как отдельные участники влияют и испытывают на себе влияние изменений в их ближайшем сетевом окружении с течением времени. DyNAM обеспечивают тонкое понимание того, как локальные взаимодействия

способствуют глобальной сетевой динамике.

**Реляционные модели событий (REM).** Реляционные событийные модели, впервые предложенные К. Баттсом [60], работают на пересечении временного моделирования и моделирования на основе связей. REM предназначены для анализа данных с временными метками, где события или взаимодействия происходят в определенные моменты времени. Они позволяют выявить временные зависимости и последовательности событий, определяющие развитие сети, что делает их подходящими для областей, где важно точное время события.

Временной стохастический подход открывает большие перспективы в различных областях, включая социальные науки, эпидемиологию, коммуникационные сети и т.д. В наукометрии, включающей количественный анализ научной литературы, коллабораций и распространения знаний, все чаще признается ценность анализа временных сетей [21, 219, 361, 404]. Временной стохастический подход представляет собой мощную призму, через которую исследователи могут изучать меняющийся ландшафт научных коммуникаций, распространения знаний и сетей сотрудничества. По мере развития наукометрии включение временных стохастических подходов расширяет аналитические возможности этой области.

### 3.11.3 Базовая модель ERGM

В данном разделе представлена общая формула экспоненциальных моделей случайных графов. Существуют различные разновидности ERGM, но суть базовой ERGM заключается в обнаружении того, как формирование и исчезновение отдельных связей влияет на сетевые конфигурации (подсети) и на глобальную структуру сети. Иными словами, базовая модель ERGM концентрирует внимание на связях между узлами.

Идея, лежащая в основе базовой модели ERGM, заключается в следующем [358]. Данна наблюдаемая сеть  $N$  с  $E$  бинарными связями (которые либо присутствуют, либо отсутствуют, но не имеют значений) и  $V$  узлами.  $\mathcal{N}$  содержит множество всех возможных конфигураций связей сети  $N$  с таким же, как в  $N$ , количеством узлов. Для оценки правильной вероятностной модели для сети  $N$ , применяется подход максимального правдоподобия. С помощью него ищется модель, которая максимизирует вероятность наблюдения исходной сети  $N$ , которую мы *действительно наблюдали*,  $\mathcal{P}(N)$ , где  $\mathcal{N}$  – это набор всех возможных сетей, которые мы могли бы наблюдать.

Ниже представлена формула вероятности наблюдения  $N$  в базовой модели ERGM:

$$\mathcal{P}(N, \theta) = \frac{\exp\{\theta' \mathbf{h}(N)\}}{\sum_{N^* \in \mathcal{N}} \exp\{\theta' \mathbf{h}(N^*)\}},$$

где -  $\theta$  вектор вещественнонзначных параметров; -  $\mathbf{h}(N)$  вектор статистик наблюдаемой сети (напр. число связей, число треугольников); -  $N^*$  – это один из элементов  $\mathcal{N}$ .

Для простоты интерпретации разобьем уравнение на четыре части: -  $\mathbf{h}(N)$  отражает статистики сети; -  $\theta$  содержит эффекты; -  $\exp\{\theta' \mathbf{h}(N)\}$  придает положительный вес наблюдаемой сети  $N$ ; -  $\sum_{N^* \in \mathcal{N}} \exp\{\theta' \mathbf{h}(N^*)\}$  нормализует все возможные конфигурации  $N$  в  $\mathcal{N}$ .

Базовая модель ERGM, как и другие разновидности, основаны на некоторых теоретических предположениях о сетях [7]:

1. Сети возникают локально.

2. На связи в сети влияют как эндогенные, так и экзогенные эффекты.
3. По сетевым характеристикам можно судить о протекающих в сети структурных процессах.
4. Несколько структурных процессов могут протекать в сети одновременно.
5. Сети, с одной стороны, структурированы, но, с другой, случайны.

### 3.11.4 Спецификация ERGM

Можно выделить три вида процессов формирования связей в сетях. Как показано на fig. 12, к ним относятся самоорганизующиеся сетевые процессы (self-organizing network processes); процессы, основанные на атрибутах акторов (attribute-based processes), и экзогенные диадические ковариаты (exogenous dyadic covariates) [7].

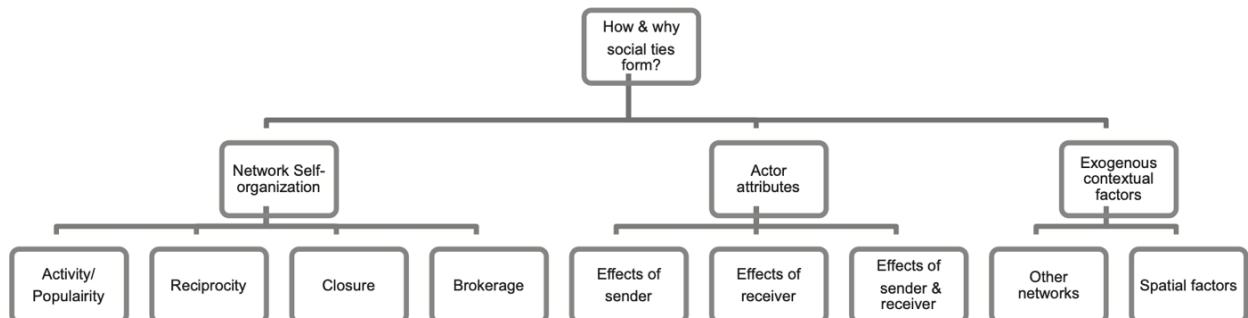


Figure 12: Классификация процессов формирования социальных связей

*Сетевая самоорганизация (Network Self-Organization).* Сетевая самоорганизация подразумевает присущую связям в сети способность самоорганизовываться в различимые паттерны под влиянием определенных типов связей. Данные эффекты называются “эндогенными”, поскольку являются результатом внутренней динамики связей сети. Можно также встретить обозначение эндогенных эффектов как “чисто структурных”, из-за отсутствия влияния атрибутов акторов или внешних факторов на связи в сети. Классическим примером служит степенной эффект (degree-based effect), широко известный в социальных науках как эффект Мэтью [314]. Данный эффект подразумевает, что чем популярнее узел в сети, тем большую популярность он приобретает.

*Атрибуты акторов (actor attributes).* Оказывать влияние на процесс формирования связей также могут различные атрибуты акторов: демографические характеристики, статус занятости, установки и т.д. В контексте ERGMs обычно используется термин “эффекты акторов-взаимодействий” (actor-relation effects), обозначающий влияние определенного атрибута актора на связь в сети. В качестве примера можно привести гомофилию – тенденцию образования связей между узлами с одинаковыми атрибутами.

*Экзогенные контекстуальные факторы: диадические ковариаты (exogenous contextual factors: dyadic covariates).* Экзогенные контекстуальные факторы часто рассматриваются как ковариаты диадической связи (то есть как влияющие на связь характеристики двух акторов), хотя ими и не ограничиваются. Например, диадическая ковариация может включать другую социальную сеть как фиксированный внешний компонент модели. В таком сценарии ERGMs может быть использован для проверки того, может ли наличие ковариационной связи предсказать возникновение соответствующей связи в интересующей нас сети. Например, рассматривая, как работники вступают в общение со своим

руководителями, ERGMs позволяют определить, как исходящие структуры с централизованными полномочиями взаимодействуют с восходящими неформальными сетями.

Как было указано выше, термины в ERGMs отличаются от тех, которые используются в традиционных статистических моделях. В обычной модели набор данных представляет собой набор переменных (результатирующей/результатирующих и предикторов), которые, хоть и могут коррелировать между собой, измеряются независимо для каждого наблюдения. Однако в ERGMs предикторы принимают особую форму – это функции, которые относятся к связям.

Список терминов из пакета ergm с краткими представлен в [277]. Данные термины определяются с использованием формулы R, которая включает как сеть, так и сетевые статистики:

$$y \sim <term1> + <term2> + \dots,$$

где  $y$  – объект сети, а  $<term1>$  и  $<term2>$  – предопределенные термины, выбранные из списка [277].

Рассмотрим наиболее распространенные термины для направленных и ненаправленных связей в пакете ergm в R.

- *Edges* – сетевая статистика, обозначающая количество связей в сети. Для ненаправленных сетей значение edges равно  $kstar(1)$ ; для направленных – как  $ostar(1)$ , так и  $istar(1)$ .
- *Density* – сетевая статистика, обозначающая плотность сети. Для ненаправленных сетей плотность равна  $kstar(1)$  или значению edges, деленному на  $n(n - 1)/2$ ; для направленных сетей плотность равна значению edges или  $istar(1)$  или  $ostar(1)$ , деленному на  $n(n - 1)$ .
- *Mutuality* – сетевая статистика (только для направленных сетей), обозначающая количество пар акторов  $i$  и  $j$ , для которых существуют  $(i \rightarrow j)$  и  $(j \rightarrow i)$ .
- *Asymmetric dyads* – сетевая статистика (только для ненаправленных сетей), обозначающая количество пар акторов, для которых существует либо  $(i \rightarrow j)$ , либо  $(j \rightarrow i)^*$ .

Возможно включить в модель эффекты атрибутов узлов (*nodal attribute effects*), то есть основные эффекты (*main effects*) и эффекты взаимодействия (*interaction effects*). Первые могут быть использованы для непрерывных ковариат или дискретных факторов; а последние – для связей, относящихся к категориальным атрибутам узлов.

Подобно узлам, атрибуты, относящиеся к диадам и связям, могут влиять на формирование связей. Атрибуты диад (*dyadic attributes*) включают тип связей (например, родство или неродство) и наличие нескольких разных связей между узлами (*multiplexity*). Атрибуты связей (*edge attributes*) включают в себя как атрибуты диад, так и специфические свойства, уникальные для связей (например, ее продолжительность).

Распределение степеней (*degree terms*) отражает частотное распределение степеней узлов, включая каждый узел только один раз. Распределение звездных конфигураций (*star terms*), напротив, отражает распределение “к-звездных” конфигураций, где один и тот же узел может присутствовать, и соответственно, подсчитываться, в нескольких конфигурациях. Для анализа доступны как параметрические линейные комбинации, так и полностью непараметрические вариации обеих статистик.

В заключение, ERGMs предоставляют возможность проверять и сравнивать в единой аналитической рамке различные гипотезы о том, что влияет на возникновение наблюданной структуры сети.

### 3.11.5 Оптимизация, оценка модели и критерий соответствия

Задача статистического вывода – согласовать распределение отдельных статистик с распределением наблюдаемой сети, по сути, подобрать модель, которая обеспечивает наиболее надежную поддержку данных. Мы устанавливаем это соответствие, определяя распределение таким образом, чтобы значения статистик из него в среднем совпадали с наблюдаемыми [lusher2013?]. Определение адекватности модели, характеризуемой вектором параметров, зависит от ее способности точно воспроизводить структурные особенности, лежащие в основе сети. По сути, оптимизация заключается в оценке того, насколько эффективно сети, полученные в результате моделирования, воспроизводят заданные структурные особенности сети. Эти структурные особенности могут включать такие показатели, как количество связей, транзитивных триад, реципрокность и др.

Важно отметить, что данная оценка относится в первую очередь к подгонке модели для конфигураций, явно включенных в модель. Однако необходимо понимать, что эта оценка не является полной оценкой соответствия (Goodness of Fit, GoF). Оценка соответствия выходит за эти рамки и включает в себя оценку того, насколько хорошо модель отражает закономерности, которые не были явно смоделированы, обеспечивая тем самым более полную оценку общей адекватности модели. Методики, используемые в процессе оценки, могут отличаться в зависимости от конкретного программного обеспечения, однако все они имеют общий подход, основанный, прежде всего, на оценке максимального правдоподобия (MLE), проводимой в рамках моделирования с использованием Марковской цепи Монте-Карло (MCMC).

В общем виде основные этапы процесса оценки включают в себя:

1. Инициализация значений параметров. Начните с получения начальных значений параметров, обычно с помощью процесса инициализации.
2. Генерация случайных графов. Приступают к генерации случайных графов при существующем векторе параметров. Эти синтетические графы генерируются в процессе моделирования.
3. Обновление значений параметров. Обновление значений параметров путем оценки распределения сгенерированных графов по сравнению с наблюдаемыми графиками.
4. Итеративное уточнение. Итерационный процесс генерации случайных графов и обновления значений параметров (шаги 2-3) выполняется до тех пор, пока не будет достигнута точка сходимости, означающая стабилизацию оценок параметров.

Такой итерационный и имитационный подход является основополагающим при оценке параметров в рамках экспоненциальных моделей случайных графов (ERGM) и им подобных моделей. Хотя у нас есть возможность проводить тесты оценки соответствия (GoF) для отдельных параметров и наборов параметров (тест Вальда, тест множителей Лагранжа и тест отношения правдоподобия), важно отметить, что эти тесты требуют спецификации конкретной альтернативной модели [7]. Следовательно, задача сводится к оценке пригодности данной модели по отношению к альтернативной, что ставит наши результаты в зависимость как от выбора модели, так и от наличия подходящих альтернатив. Для устранения этого недостатка Робинс, Паттисон и Вулкок [326] предложили подход имитационного моделирования. Он позволяет исследовать целый спектр характеристик графа. Основная концепция заключается в оценке способности модели эффективно отражать те аспекты данных, которые не были явно заложены в саму модель. Например, может ли параметр ребра и чередующиеся треугольники

адекватно объяснить наблюдаемую среднюю длину пути или наблюдаемое распределение степеней? Такая процедура позволяет провести более комплексную оценку эффективности модели, чем обычная проверка гипотез.

### 3.11.6 Темпоральный ERGM (TERGM)

Как определено в работе Лейфельда и Кранмера, TERGM развиваются идею, заложенную в ERGM [231]. Они определяют вероятность сети на текущем временном шаге  $t$  как функцию не только суммы подсчетов подграфов текущей сети, но и предыдущих сетей до временного шага  $t - K$ :

$$P(N^t | N^{t-K}, \dots, N^{t-1}, \theta) = \frac{\exp(\theta^T h(N^T, N^{t-1}, \dots, N^{t-K}))}{c(\theta, N^{t-K}, \dots, N^{t-1})}.$$

При этом предполагается, что статистические показатели, полученные на основе связей между временем  $t - K$  и временем  $t$ , эффективно отражают присущие сети зависимости в момент времени  $t$ . Эта простая идея лежит в основе TERGM. В знаменатель этой формулы входит нормирующая константа, аналогичная той, что используется в ERGM. На следующем этапе определяется вероятность, связанная с временным рядом сетей, путем вычисления произведения всех временных периодов:

$$P(N^{K+1}, \dots, N^T | N^1, \dots, N^k, \theta) = \prod_{t=K+1}^T P(N^t | N^{t-K}, \dots, N^{t-1}, \theta).$$

Это представляет собой простое расширение ERGM на последовательность сетей. Для учета временных зависимостей между последовательными временными шагами вводится статистика сети  $h$ , позволяющая включать в анализ временной аспект. Лейфельд и коллеги предлагают исчерпывающее рассмотрение этого вопроса [231].

Граф зависимости TERGM, формально определяющий зависимость одной диады от другой, может моделировать зависимость между моделируемыми переменными в нескольких различных временных точках [231]. В отличие от многих других моделей, TERGM воздерживается от предположений об интервалах между последовательными временными шагами, будь они длинными или короткими, непрерывными или дискретными. Она не зависит от того, последовательно или одновременно формируются ребра сети в процессе генерации данных. Основным требованием является то, что результат может быть переведен в термин зависимости, который легко вписывается в вектор  $h$ . Эта гибкость, присущая TERGM, коренится в некоторых ограничениях, накладываемых на статистику  $h$ , что позволяет ей учитывать широкий спектр сетевых структур.

Оцениваемые параметры можно рассматривать как логарифмические коэффициенты вероятности установления связи внутри сети с учетом конфигурации и выбранных параметров остальной части сети и влияния до  $K$  предшествующих сетей. Для оценки параметров часто используется оценка максимального правдоподобия Марковской цепи Монте-Карло (MCMC-MLE) [154]. Эти методы оценки удобно реализованы в пакете *btergm*, специально разработанном для среды статистических вычислений *R* [231].

### **3.11.7 Исследование структур в научометрических исследованиях: перспективы и использование ERGM и TERGM**

Применение ERGM и TERGM к библиометрическим сетям выглядит очень логичным, и исследование в этой сфере появились не в последние годы. В тексте ниже мы показываем какие исследования с применением экспоненциальных случайных графов в сфере библиометрических исследований существуют, какие научные сообщества в каких странах и регионах были исследованы, и какие выводы могут быть сделаны относительно библиометрических сетей. ERGM позволяет моделировать не отдельные отношения между акторами, а целую сеть; однако, экспоненциальные модели случайных графов могут работать только с бинарными данными и не адаптированы для динамического анализа. TERGM является продолжением экспоненциальных моделей случайных графов – темпоральным экспоненциальным моделированием случайных графов. Это разновидность модели, рассматривающей отдельные состояния графа в равноудаленные моменты времени. В случае библиометрических исследований с помощью ERGM можно рассматривать библиометрические сети для одного момента времени, а с помощью TERGM можно моделировать ту же библиометрическую сеть для разных лет и, соответственно, оценивать эффекты, влияющие на сети в динамике. Однако TERGM также может работать только с бинарными данными, и у исследователей возникают вопросы по интерпретации влияния временных зависимостей на уровне сетевых связей [49].

**3.11.7.1 Библиометрический анализ: применение ERGM** Модель ERGM появилась около 15 лет назад. Распространению моделей ERGM во многом способствовало появление пакета statnet и реализация в нем ERGM [215, 216].

Большинство статей, использующих ERGM для моделирования библиометрических сетей, представляют работы, анализирующие состояние той или иной научной области, и авторы статей являются представителями этой дисциплины. Например, Окамото Джанет – директор Центра оценки здоровья населения – в 2015 году опубликовала статью, в которой проанализировала сеть партнерств в области изучения неравенства в здравоохранении [384]. Аналогичная ситуация складывается с исследователями в области компьютерных наук [23], информационно-поисковой сферы [435] или исследователями научных инноваций из региона на западе Китая [180]. Хотя эти исследования демонстрируют лишь практическую реализацию модели ERGM и каждое из них содержит ограничения, реалистичные для всех моделей ERGM, они могут продемонстрировать важные наблюдения о своих предметных областях и быть полезными для ученых, академических институтов, государств и бизнеса.

Другой уровень исследований представляет собой изучение научных коллабораций на уровне отдельных стран или регионов. Заслуживают внимания следующие исследования: изучение сети патентного цитирования в Европе [66], изучение словенских научных сообществ на примере 4 наук (физики, математики, биотехнологии и социологии) [218], исследование сотрудничества британских исследователей по финансируемым проектам [360]. Хотя на данном этапе научные исследования могут сказать больше о состоянии науки и структуре научных коллабораций в стране или регионе, они страдают и от более серьезных ограничений: например, авторы исследования европейских патентов говорят о том, что за определенными ссылками на патенты в некоторых компаниях может стоять структура и цель, которые они не могут четко определить [66].

Последний уровень библиометрических исследований с использованием ERGM – это проекты, в которых исследователи пытаются понять общий характер сотрудничества между различными дисциплинами или субдисциплинами или состояние международной науки. Например, в 2013 году, в момент начала широкого распространения ERGM, Даниэле Фанелли решил выяснить, как выглядят 12 научных дисциплин: как объясняет Фанелли, на структуру дисциплин влияет характер дисциплины: для сложных и специфических явлений исследователи с меньшей вероятностью достигнут теоретического и методологического консенсуса [118].

**3.11.7.2 Библиометрический анализ: применение TERGM и VERGM** Помимо ERGM, ученые могут использовать также TERGM – Temporal Exponential Random Graph. Исследований с применением TERGM в библиометрическом анализе не так много, и это связано с характером моделей: они требовательны к вычислениям. В 2023 году Тревис Уитшелл решил узнать, влияет ли политический режим государств на международное научное сотрудничество, проанализировал данные о международном научном сотрудничестве по 170 странам за 2008-2017 гг. и обнаружил, что демократический режим является хорошим предиктором более частого международного научного сотрудничества [419]. Это исследование является хорошим примером исследования, охватывающего сразу большой временной период и использующего TERGM. Кроме того, Уитшелл использует модель Value-temporal Exponential Random Graph (VERGM), и именно ее использование и оценка вероятности возникновения новых отношений на небинарных данных о сотрудничестве дает Уитшеллу наиболее точные результаты.

Как видно, авторы работают с различными сетями: сетями соавторства, сетями патентного цитирования, бимодальными сетями научного сотрудничества и финансирования исследовательских проектов. В исследованиях в основном используется ERGM, однако другие исследователи применяют TERGM и VERGM. В целом область исследований на стыке библиометрических исследований и применения ERGM можно описать как состоящую из трех видов исследований: исследований отдельной области научного знания, исследований сетей отдельных дисциплин, исследований состояния науки в целом.

### **3.11.8 Заключение**

В заключение следует отметить, что область научного взаимодействия и сетей сотрудничества испытывает острую потребность в разработке моделей, основанных на данных, для лучшего понимания процесса распространения знаний. Изучение научного взаимодействия, социальных и когнитивных структур в различных научных областях успешно проводится с помощью библиометрии и наукометрии, причем особое внимание уделяется анализу временных библиографических сетей, таких как соавторство, цитирование, совместное цитирование и библиографическое сопряжение. Эти модели должны включать в себя как реалистичные временные структуры, так и кросс-секционные особенности.

Исследователи применяют различные методы для изучения научного сотрудничества, при этом анализ сетей соавторства является одним из доминирующих подходов благодаря простоте извлечения данных из баз данных публикаций. Однако он требует тщательной очистки данных.

В докладе обсуждались возможности применения экспоненциальных моделей случайных графов (ERGM) в научометрических исследованиях, подчеркивалась значимость стохастических методов в

сетевом анализе. Детерминированный подход служит основой для более сложных аналитических методик и включает в себя глобальные свойства, локальные свойства и методы разбиения для анализа сетей. Также было рассмотрено применение ERGM и TERGM к библиометрическим сетям, что свидетельствует об их универсальности при моделировании различных типов сетей, таких как сети соавторства, сети цитирования патентов, сети финансирования научных проектов.

В целом, пересечение библиометрических исследований и применения ERGM охватывает исследования конкретных областей научного знания, изучение сложных сетей и анализ состояния самой науки. Это направление исследований подчеркивает потенциал анализа временных сетей, позволяющий пролить свет на меняющийся ландшафт научной коммуникации и генерации знаний в области научометрии.

### **3.12 Современные методы анализа неструктурированной текстовой информации: систематизация и сравнительный анализ**

#### **3.12.1 Введение**

Цель настоящей главы и приведенного в ней аналитического обзора состоит в систематизации различных современных подходов к текстовому анализу на основе нейросетевых моделей и их сравнительном анализе для оценки эффективности применения для прикладных задач. Достижение этой цели требует решения следующих задач:

1. Рассмотреть ключевые прикладные задачи текстового анализа.
2. Проанализировать энкодер и декодер модели

Современные методы анализа неструктурированной информации решают прикладные задачи обработки текстов, включающие в себя следующие задачи:

- Извлечение именованных сущностей (Named Entity Recognition, NER). Это задача выделения и классификации именованных сущностей, таких как имена, места, даты, организации и другие в тексте.
- Извлечение отношений (Relation Extraction). Задача определения связей между различными именованными сущностями в тексте, например, определение связи между человеком и местом работы.
- Перефразирование (Paraphrase Generation). Генерация семантически эквивалентных предложений или выражений с целью переформулирования их с сохранением смысла.
- Суммаризация (Summarization). Автоматическое создание кратких сводок из больших объемов текста. Суммаризация может выполняться как по одному тексту, так и по массиву текстов.
- Ответы на вопросы (Question Answering, Q&A). Построение моделей, способных отвечать на вопросы, заданные в естественном языке.
- Классификация текста (Text Classification). Разделение текста на категории или классы на основе его содержания или тематики.
- Кластеризация текста (Text Clustering). Структурирование массива текстовых документов на основе семантической близости их содержания без знания изначальной структуры массива.

- Определение частей речи (Part-of-Speech Tagging, POS). Присвоение токенам в тексте соответствующих частей речи, таких как существительные, глаголы, прилагательные и др.
- Машинный перевод (Machine Translation). Автоматический перевод текста с одного языка на другой с сохранением смысла.
- Разрешение семантических ссылок (Coreference Resolution). Определение, к каким сущностям или объектам в тексте относятся местоимения или фразы.

Современные модели анализа неструктурированной информации базируются на представленной в 2017 году статье “Attention is All You Need” [401]. Эта архитектура нашла широкое применение благодаря своей способности эффективно работать с последовательностями большой длины без использования рекуррентных или свёрточных слоев.

### 3.12.2 Основные компоненты модели трансформер

1. Encoder-Decoder архитектура Трансформер состоит из двух основных частей: энкодера и декодера. Энкодер принимает на вход последовательность слов и преобразует её в скрытые представления. Декодер занимается генерацией текста, используя эти скрытые представления энкодера для предсказания следующих слов.
2. Механизм внимания (Self-Attention) Одной из ключевых концепций трансформера является механизм внимания, позволяющий модели эффективно работать с контекстом. Этот механизм позволяет модели “фокусироваться” на различных частях входной последовательности при генерации выходных предсказаний. Self-Attention позволяет модели учитывать дальние зависимости в тексте, что делает её способной лучше понимать контекст.
3. Позиционное кодирование Трансформер обрабатывает всю последовательность токенов целиком, в отличие от рекуррентных сетей, обрабатывающих последовательность токенов один за другим. Для сохранения информации о позициях слов в последовательности в трансформере используется позиционная кодировка. Это позволяет модели учитывать порядок слов при реализации механизма внимания.
4. Многоуровневые слои Трансформер состоит из множества слоев энкодера и декодера. Каждый слой состоит из нескольких подслоёв: механизм внимания, полносвязные нейронные сети и нормализация. Многие такие слои обеспечивают более глубокое представление для текста.

Благодаря возможности обрабатывать всю последовательность целиком Трансформер позволяет эффективно значительно повысить количество текстов, обрабатываемое за единицу времени, что ускоряет процесс обучения. Благодаря механизму внимания трансформер лучше улавливает дальние зависимости в тексте.

Модель трансформер показывает высокую гибкость и применима к различным задачам NLP без необходимости переобучения модели с нуля. На основе модели трансформер сделано большинство современных языковых моделей, разделившихся на генеративные и энкодерные. Эти модели стали краеугольными камнями в различных задачах, таких как классификация текстов, извлечение сущностей, определение тональности и кластеризация данных. ### Генеративные модели Генеративные модели,

такие как рекуррентные нейронные сети (RNN) и свёрточные нейронные сети (CNN), создают данные, имитируя вероятностные распределения слов или символов в тексте. Примерами таких моделей являются LSTM (Long Short-Term Memory) и GRU (Gated Recurrent Unit), способные учитывать контекст и последовательность слов при генерации текста. Однако, генеративные модели могут страдать от проблемы затухающего или взрывающегося градиента и могут иметь ограниченные возможности в генерации длинных текстов.

### 3.12.3 Энкодерные модели

Энкодерные модели, такие как Transformer и его вариации (например, BERT, GPT), работают как механизмы кодирования и понимания текста. Они преобразуют входные данные в векторные представления с помощью механизмов внимания и многоуровневых архитектур. Энкодерные модели, обученные на больших объемах данных, обычно демонстрируют высокую производительность в решении задач NLP.

С непрерывным развитием технологий глубокого обучения в последние годы, предварительно обученные языковые модели Pre-trained Language Models (PLM), которые обучаются с использованием самообучения на огромных корпусах текста, широко используются в области обработки естественного языка (NLP) и показывают лучшие результаты в различных задачах обработки текста. В отличие от традиционного обучения с учителем, PLM, основанные на самообучении, предварительно обучаются на обширных неразмеченных текстах, а затем дообучаются для решения конкретных задач на маломасштабных размеченных данных.

### 3.12.4 Сравнение моделей для задач NLP

**Классификация текстов** При выполнении задачи классификации текстов, где необходимо отнести текст к определенной категории, энкодерные модели, такие как BERT, показывают высокую точность благодаря способности улавливать семантическую информацию и контекст предложений.

**3.12.4.1 Извлечение сущностей** В задаче извлечения сущностей, где требуется определить и классифицировать именованные сущности в тексте (такие как имена, места, даты и т.д.), энкодерные модели, такие как BERT и GPT, успешно применяют механизмы внимания для извлечения информации из текста.

**3.12.4.2 Определение тональности** Для определения тональности текста, где необходимо понять отношение текста к определенной эмоциональной окраске (позитивной, негативной, нейтральной), как генеративные, так и энкодерные модели могут показать высокую точность. Однако, модели, основанные на энкодерах, обычно имеют более широкий спектр понимания контекста, что делает их более эффективными в данной задаче.

**3.12.4.3 Кластеризация данных** В задаче кластеризации, где необходимо группировать похожие текстовые данные, энкодерные модели, такие как BERT, имеют тенденцию лучше улавливать семантические связи и создавать более качественные кластеры.

### **3.12.5 Выводы**

Генеративные и энкодерные языковые модели имеют свои преимущества и недостатки в решении различных задач NLP. Выбор модели зависит от конкретной задачи, доступных ресурсов и требуемой точности. В последнее время энкодерные модели, особенно на основе трансформеров, демонстрируют значительные успехи благодаря своей способности к обучению на больших корпусах текста и высокой гибкости в различных задачах NLP.

## **3.13 Внедрение внешних знаний в языковые модели**

Большие языковые модели, основанные на BERT, GPT, T5, постоянно обновляют рекорды во многих задачах понимания и генерации естественного языка (NLU и NLG). Однако, из-за отсутствия механизмов представления знаний, большие языковые модели имеют ограниченную возможность обучения которая может быть восполнена благодаря использованию механизмов внедрения знаний.

Знания, внедряемые в языковую модель могут иметь разную структуру и происхождение. Так, в работе [181], авторы выявляют четыре крупных категории внедряемой информации в зависимости от её происхождения: - использование лингвистических знаний; - использование текстовых знаний; - использование графа знаний; - использование правил и эвристик.

### **3.13.1 Использование лингвистических знаний**

Лингвистические знания, в основном представленные в виде лексической информации и синтаксических деревьев, являются наиболее распространенной вспомогательной информацией для языковых модели [430]. Лингвистические знания также включают информацию о частях речи (POS-тегирование) и тональность слов [225]. Модель LIBERT (lexically-informed BERT) добавляет к традиционным задачам для предобучения BERT прогнозирование лексических отношений. Модель прогнозирует отношения синонимии, гиперонимии и гипонимии, что позволяет в дальнейшем лучше моделировать семантическую информацию.

SenseBERT [234] добавляет к стандартной задаче заполнения замаскированного токена прогнозирование таксономический класс токена, например, существительное.еда, существительное.состояние за счет чего лучше учитывает семантический контекст.

SKEP [394] схожим образом кодирует тональность слов что позволяет повысить внимание модели к тональной лексике и дает прирост в интерпретабельности модели.

DictBERT [68] принимает лексические знания из словаря в качестве внешнего источника и повышает качество предобучения за счет контрастного обучения.

Интерес представляет модель Syntax-augmented BERT [337] использующая графовые нейронные сети для моделирования синтаксических связей, полученных из деревьев зависимостей, для улучшения языкового моделирования. Эта модель демонстрирует подход по встраиванию выхода GAT в языковую модель

### **3.13.2 Использование текстовых знаний**

Текстовые знания обычно используются для повышения качества поиска или вопрос-ответных систем. UDT-QA [253] используют текст, граф знаний и таблицы знаний вместе, и встраивают объединенный

вектор в verbalizer-retriever-reader цепочку.

KNN-LM [198] выбирает ближайших  $K$  ближайших соседей из обучающих образцов в языковую модель в качестве гипотез чтобы повысить качество прогнозирования маскированных токенов. ExpBERT [281] [46] и KEAR [426] также включают текстовые описания в свои модели для улучшения работы механизма внимания. Kformer [429] [48] получает некоторые внешние текстовые знания через извлечение и вводит их в слой полносвязанной сети.

### 3.13.3 Использование графа знаний

Граф знаний можно описать как набор троек формата  $, ,$ , где каждая тройка отражает отношение между двумя сущностями через определенное отношение [189]. KG может быть представлен как  $\boxtimes = \{\boxtimes, \boxtimes, \boxtimes\}$ , где  $\boxtimes$  - множество сущностей,  $\boxtimes$  - множество отношений и  $\boxtimes$  - множество троек. Термин RDF - это либо URI и  $\boxtimes U$ , либо пустой узел  $b \boxtimes B$ , либо литерал  $l \boxtimes L$ . Узлы URI (или IRI) служат для глобальной идентификации сущностей в Web, а узлы литералов - для их идентификации.

Благодаря более структурированной по сравнению с текстом информации граф знаний может быть более применимым для обучения моделей чем предыдущие источники [74] [57], [119] [58], [235].

В общем случае архитектура встраивания графов знаний в языковую модель показана на рисунке ниже.

Один из подходов к обучению модели является разработка механизма предварительного обучения, с использованием триплетов. [316], [410], [437], [405]. ERICA [316] представляет подход на основании контрастного обучения для дискриминации сущностей, так и отношений. KEPLER [410] обучает функцию потерь на триплетах и языковую модель одновременно.

DKPLM [437] обогащает семантическую информацию редко встречающихся сущностей знаниями из графов знаний. KP-PLM [405] разрабатывает два механизма предобучения, связанные с знаниями о тройках, чтобы интегрировать знания троек в несколько непрерывных подсказок для задач естественного языка.

Второй способ - изменить механизм внимания модели [379], [240]. K-BERT [240] использует уровень знаний для инъекции соответствующих троек из графа знаний во входное предложение и преобразует его в дерево предложения с обилием знания для управления областью каждого слова в предложении, предотвращая отклонение предложения от правильной семантики. Если K-BERT расширяет входной текст в дерево предложения, то базовой концепцией CoLAKE [379] является расширение контекста ввода в графы слов-знаний (WK-графы), а затем подача этих построенных WK-графов в маскированное внимание, чтобы собирать информацию узлов.

Третий способ - изменить структуру модели, который обычно включает модуль графа знаний. K adapter [408] и KB-adapters [112] интегрируют знания в PLM через внешние адаптерные модули. KLMO [161] использует компонент, названный агрегатор знаний, для слияния вложений входного текста и КГ, в котором применяется перекрестное внимание на уровне сущности КГ для интерактивного моделирования сегментов сущностей в тексте, а также сущностей и отношений в КГ. KERM [104] разрабатывает модуль внедрения знаний, который объединяет информацию между корпусом текстов и КГ для задачи переупорядочивания прохода, как показано на рисунке ниже. JointLK [380] и GreaseLM [439] используют GNN для моделирования извлеченных графов знаний и связывают LM с модулями GNN для совместного

рассуждения при рассуждении о здравом смысле.

#### 3.13.4 Использование правил и эвристик

Логические правила всегда содержат четкие предикативные инструкции и могут формализовывать знания из внешних источников [335]. Включение этого типа знаний в большие языковые модели может повысить обучаемость благодаря высокой интерпретируемости. Так, RuleBERT [338] использует полученные правила Хорна из существующего корпуса, чтобы создать обучающий набор данных и затем донастроить модель на нем. Он применяет вероятностную модель ответа для извлечения мягких правил из языковой модели. Результаты показывают, что языковые модели, работающие с мягкими правилами на естественном языке, могут повысить свою производительность для задач умозаключения. Кроме того, PTR [152] может включать логические правила для составления вручную доменно- и задачно- специфических подсказок, позволяющих модели закодировать задачно-связанные предварительные знания при настройке подсказок и генерировать более интерпретируемые подсказки. И RuleBERT, и PTR включают правила знания на этапе донастройки модели.

#### 3.13.5 Использование социально-сетевых графов

Одним из источников знаний для улучшения качества модели является социальный граф, построенный под решаемую задачу. Эта информация в большом количестве доступна в онлайн социальных сетях.

При этом группы онлайн-социальных сетей являются предпочтительными для моделирования структуры языка из-за следующих характеристик: - Группы и публичные страницы (далее “группы”) имеют свои собственные страницы. Тексты, размещенные на страницах групп, в основном монотематичны, так как пользователи групп делятся общим интересом или обсуждают новости, важные для определенного географического региона. В обоих случаях возможно выделить специфическую лексику и речевые особенности группы. - Количество групп значительно меньше, чем количество пользователей. Это позволяет создать языковую модель, подходящую для всей онлайн-социальной сети, без значительных затрат на фильтрацию узлов и вычислений. - Группы генерируют основную часть текстового контента, в то время как многие пользователи социальных сетей не пишут ни одного слова в течение нескольких лет, так как выступают только в роли потребителей контента.

В то же время интересы пользователей довольно легко выражаются через группы, на которые они подписаны. Рассматривая группу как автора текстов, написанных от имени этой группы можно оценить близость групп через пересечение общих пользователей, чтобы группы с совпадающими подписчиками имели схожие языковые модели.

Предлагаемая модель учитывает характеристики домена, используя предварительно вычисленный социальный вектор для анализа каждого токена входного текста. Общий процесс обучения следующий:

- Генерация матриц смежности на основе сетевых данных - подготовка матриц для оценки смежности двух групп на основе взаимных членов группы.
- Обучение социальных векторов - получение векторов авторов с использованием алгоритмов факторизации и случайного блуждания.

Обучение BERT с использованием предварительно обученных социальных векторов.

**3.13.5.1 Генерация векторных представлений групп** При вычислении социального вектора целесообразно использовать информацию о локальной среде сообщества, а также описания его глобального положения относительно всех групп. Для имитации локального контекста хорошо подходит алгоритм DeepWalk. Чтобы учесть структуру наших социальных графов на более глобальном уровне, можно также использовали факторизацию различных видов попарных матриц расстояний между группами.

Альтернативой подходу выше является факторизация попарной корреляционной матрицы групп. Рассматривя группу как вектор нулей и единиц длиной, равной общему числу пользователей в социальной сети  $N$ , и содержащий 1 для пользователей, подписанных на моделируемую группу, и 0 в противном случае. Вычисление корреляции может быть реализовано без загрузки всей матрицы в оперативную память на основе легко вычисляемых переменных: для набора A подписчиков группы a и набора B подписчиков группы b мы, как описано в предыдущем разделе, вычисляем размер пересечения  $|A \otimes B|$ .

**3.13.5.2 Обучение модели BERT** Полученные векторы, полученные в результате случайного блуждания или SVD, могут интегрированы в существующую модель BERT для того чтобы научить модель обращать внимание на вектор социальной сети. Для этого может использоваться несколько различных способов встраивания социальной информации:

- добавление специального вектора социальной сети, который объединяет оба характеристики в начале каждой последовательности (внедрение токена Zero).
- добавление специального слоя “социального внимания” (SAT) на различных позициях в существующей модели BERT, как описано ниже.

Общая схема обеих подходов показана на рисунке ниже.

Для лучшего внедрения информации о социальной сети в языковую модель предлагается использовать специальный слой SAT, показанный на рисунке ниже.

Механизм встраивания зависит от двух гиперпараметров:  $i$  - номер слоя BERT, выбранный для замены на SAT слой, и  $C$  - количество каналов для использования в SAT слое. Для встраивания слоя социального внимания сначала мы предварительно обучаем базовую модель BERT на всем обучающем наборе данных в течение одной эпохи. Затем выполняется заморозка всех слоев модели и замена  $i$ -го слоя SAT слоем.

При построении SAT слоя предлагается использовать двухслойный персепtron с функцией активации GELU между слоями и активацией SoftMax после второго слоя. Векторы социальной сети передаются через этот MLP, в результате получаются новые векторы  $W$  с уменьшенной размерностью до  $C$ . Затем создается  $C$  параллельных слоев BERT, каждый инициализированный как замена  $i$ -го слоя оригинального BERT. Выход SAT слоя является суммой векторов, полученных умножением каждого из параллельных слоев BERT на соответствующий элемент социального вектора  $W$ . Идея этого метода заключается в том, чтобы обучить каждый из наших  $C$  слоев BERT настраиваться на надмножество тем социальных сетей, а затем представить каждого автора как композицию этих надмножеств.

Предложенная модель может быть оценена с использованием качества прогнозирования отсутствующего элемента в предложении и показателя перплексии, используемого в оригинальных

работах, таких как BERT и RoBERTa. Абсолютное значение перплексии для данной модели зависит от многих параметров, таких как размер словаря модели, параметры токенизации и набор данных для донастройки. Таким образом, довольно сложно оценить прямое влияние перплексии на решение какой-либо прикладной задачи. Наш случай дополнительно осложнен необходимостью подготовки собственной эталонной выборки, так как, насколько нам известно, ни один из существующих наборов данных не содержит информацию о социальных связях автора, используемых нашей моделью.

С другой стороны, разница в перплексии для одной и той же базовой модели, обученной на полностью идентичном корпусе с одинаковой предварительной обработкой, должна отражаться на качестве применяемых задач, как показано авторами статьи о RoBERTa и оригинальным BERT: по мере уменьшения перплексии, повышается качество классификации на наборе данных SST-2 (для RoBERTa и BERT), MNLI-m и MRPC (для BERT). Таким образом, разница в перплексии для двух изначально идентичных моделей BERT, обученных на одних и тех же текстах, указывает на более высокую обучаемость и дальнейшую эффективность модели с более низкой перплексией.

В оригинальной статье BERT сообщается о перплексии 3,23 для модели с 24 слоями и входного токена 1024. Модель BERT Base, обученная на том же корпусе, имеет перплексию не менее 3,99 как для английского, так и для русского языка. Это можно объяснить значительным разнообразием тем, и даже языков, охватываемых этими моделями. Поскольку тексты онлайн-социальных сетей (ОСС) являются подмножеством всего массива текста, обучение только на ОСС позволяет снизить перплексию до 2,83 для многоязыковой базовой модели BERT (RuBert OCC). Дальнейшее улучшение возможно за счет использования дополнительной информации о социальных векторах, что позволяет снизить показатель оценки до 2,72.

Такая модель может быть полезной для всех задач понимания языка, прежде всего, для сопоставление сущностей, проверка правописания и извлечение фактов. Модель показывает очень многообещающие результаты на коротких сообщениях и текстах с бедным контекстом:

- Полученные примеры демонстрируют, что модель успешно учитывает региональные особенности. Например, для шаблона “набережная [MASK]”, базовая модель BERT рекомендует “осенняя набережная”, тогда как модель, инициализированная региональными группами Санкт-Петербурга, предлагает “Невская набережная” на основе реки Невы в Санкт-Петербурге.
- Модель может быть полезна для задач прогнозирования связей на коротких текстах. Например, для шаблона “мы сегодня прочитали [MASK] Александра”, базовая модель BERT возвращает “мы сегодня прочитали Александра Королёва” (актер и продюсер), тогда как модель с инициализацией векторов группы поэзии возвращает “мы сегодня прочитали Александра Блока” (известный поэт).
- Модель также полезна при определении профессионального жаргона. Например, для заданного шаблона “Big [MASK]”, базовая модель BERT возвращает “Big bro”, тогда как модель с вектором группы Data Science возвращает “Big data”.

## 4 Возможности развития методики когнитивного интервью

Марина Буракова (КСА, ФСН), Иван Климов

## **4.1 Введение**

Стандартом социологических исследований считается проведение «пилотажа» - тестирования инструментария перед запуском «полевого» этапа. Однако на практике эта процедура реализуется нечасто: редко когда в публикациях и отчетах можно встретить хотя бы упоминание о действиях на предварительном этапе. Да и если посмотреть издающиеся учебники, пилотаж описывается не очень подробно.

«В многоплановых аналитических исследованиях, где применяется сложный методический инструментарий, может возникнуть потребность в двух или трех пилотажах. Их проведение целесообразно поручить разработчикам опросника, которые способны уловить мельчайшие недостатки тех или иных вопросов, обнаружить психологические барьеры их восприятия. В процессе пилотажа фиксируются реакция респондента на каждый вопрос, его замечания. Если он затрудняется ответить, нужно попытаться выяснить причину. В случае непонимания какого-то вопроса, последний формулируется по-новому» [453].

«Пилотажные исследования могут проводиться в два этапа: предварительная проба и генеральный пилотаж. На первом проверяются отдельные элементы методики. На втором проверяется вся процедура сбора информации. Это «разведка боем», цель которой — проверка всей полевой процедуры» [481] [Тавокин, 2009, С.11].

Существует множество способов предварительного тестирования вопросов анкеты: экспертные оценки, опрос интервьюеров, фокус-группы, эксперименты с альтернативными вариантами формулировок [272]. Начиная с 1980 годов одними из самых широко распространенных становятся методы когнитивного интервью. Когнитивное интервьюирование выступает в качестве одной из ключевых тем Конференции разработки, оценки и тестирования дизайна анкеты в 2002 и 2016 годах [13]. Бюро переписи населения США (Census Bureau), Национальный центр статистики здравоохранения США (National Center for Health Statistics), в России - ФОМ, ВЦИОМ и многие другие организации включают процедуру когнитивного интервьюирования в стандартный процесс разработки опросного инструментария [43]. Проблема, тем не менее, заключается в том, что в отчетах по исследованиям эти процедуры и их результаты описываются редко, что приводит к стагнации в развитии метода и подхода в секторе русскоязычной исследовательской литературы. Список работ о практике проведения когнитивного интервью немногочисленный. Основным источником, на который ссылаются большинство авторов, является работа Д. М. Рогозина [475]. Очевидна необходимость обзора основных теоретических парадигм когнитивного интервью, сравнения методологических преимуществ и ограничений каждого подхода, оценка возможностей КИ для задач не только тестирования опросного инструмента, но также и для других распространенных исследовательских задач, а также ознакомление российских исследователей с практикой проведения КИ.

## **4.2 Цели и задачи исследования**

Цель: тематизировать основные направления в использовании и развитии методики КИ, оценить эвристические возможности КИ в некоторых актуальных направлениях методических исследований.

Задачи исследования: 1. На основе анализа отечественной и зарубежной литературы определить возможности и методические ограничения основных теоретических подходов проведения когнитивного

интервью. 2. Определить возможности и ограничения методологии КИ для задачи культурной адаптации измерительных методик, определить возможности для корректировки метода когнитивного интервью применительно к задаче культурной адаптации опросного инструментария. 3. Описать ограничения и возможности работы с результатами когнитивного интервью на основе сетевого анализа и работы с неструктурированными массивами текстовых данных.

### **4.3 Практическая значимость исследования**

Практическая значимость работы – в ее нацеленности на практические потребности исследователей, занимающихся разработкой опросного инструмента и работающих с переводными методиками. Мы не только провели систематизацию подходов и методических разработок в области когнитивного интервью. Также мы провели серию экспериментов по использованию этого подхода для задач кросс-культурной адаптации опросных методик (на примере одного психометрического теста). Наш следующий шаг – разобраться в вопросе о том, как анализировать результаты КИ и провести апробацию модели анализа неструктурированной текстовой информации на основе сетевого анализа.

### **4.4 Обзор литературы**

Описание research gap: Использование в исследовательской практике переводного инструментария – довольно распространенная практика. Хорошо известны приемы, которыми пользуются исследователи для валидизации перевода (прямой и обратный перевод, перевод несколькими переводчиками с разной областью экспертизы, обсуждение разногласий и др.). Также существует обширная традиция кросс-культурной адаптации инструментария. Однако целенаправленное использование методики когнитивного интервью для этих задач встречается довольно редко. На наш взгляд, эвристические возможности КИ для задач кросс-культурной адаптации опросников остаются недооцененными и не проработанными. Для решения этой задачи мы выбрали одну из методик (психометрический тест), не существующих пока на российском рынке, но к которой у отечественных исследователей есть и интерес, и готовность ее перевести.

Степень научной разработанности проблемы. Использованные в нашей работе источники можно разделить на три группы: психометрические тесты и характеристики их эффективности; методы адаптации психометрических тестов, а также методы когнитивного интервьюирования.

1. Работы из первой группы послужили основой для составления общей картины исследования и разработки психометрических тестов. Наиболее значимыми в области психометрического тестирования являются работа Анастази “Evolving concepts of test validation” [26] и Американской психологической ассоциации (APA) “Standards for educational and psychological testing” [32].

Анастази принадлежит наиболее распространенное определение понятия психометрического теста, которое согласуется с фундаментальной в образовательном и психологическом тестировании работой “Стандарты образовательного и психологического тестирования” и включает в себя его ключевые компоненты. Анастази определяет психометрический тест как объективное и стандартизированное измерение образцов (проб) поведения.

Кроме того, изучением особенностей психометрических тестов, разработанных для задачи измерения компетенций занимались Кронбах и Мил; Хэмблтон, Сваминатан и Роджерс; Радж и Элис;

Гизелли и Браун; Гион; Моссхолдер и Арви; Бранник и Левин; Щербаум; Маккорник, Де Ниси и Шоу; Левинджер; Кларк и Уотсон; Кэмпбелл; Мэссик; Лорд и Новик.

Ключевой работой, посвященной изучению моделей компетенций личности, выступает работа Бояцис “The competent manager: A model for effective performance” [56]. В рамках данной работы было сформулировано определение компетенций, выступающее фундаментом изучаемого нами психометрического теста.

2. Разработка и тестирование подходов к адаптации психометрических тестов были описаны в руководстве Международной тестовой комиссии, работах Битон, Бомбардиер, Гуллемин и Ферраз, Солано-Флореса, БекхоФфа, Эпстайна, Санто, Фокс-Рушби, Хердмана. Ключевой работой в данном списке является руководство, разработанное Битоном, Бомбардиером, Гуллемином и Ферразом, в рамках которой нами были рассмотрены основные возможности и ограничения применения каждого из описанных авторами методов адаптации.

Одним из первых официальных документов, в котором были собраны рекомендации по переводу и адаптации психометрических тестов является сборник Международной тестовой комиссии (International Test Commission), работа над первым изданием которого была начата в 1992 году [81]. На сегодняшний день последнее издание сборника, которое разрабатывалось в период с 2005 по 2015 год, датировано 2017 годом и включает в себя шесть разделов [81]. В рамках последнего издания Международной тестовой комиссии адаптация определяется как совокупность таких процессов, как принятие решения о возможности измерения конструкта в целевой культуре, к которой адаптируется тест; подбор переводчиков; разработка схемы оценивания работы переводчиков; изменение формата проведения теста при необходимости; перевод; проверка эквивалентности переведенного варианта теста исходному [81]р.6-7.

Битон, Бомбардиер, Гуллемин и Ферраз определяют кросс-культурную адаптацию как процесс, в ходе которого исследователи преследуют цель обеспечения эквивалентности между адаптируемым и адаптированным вариантами психометрического теста на основе его содержания [42]. Исследователи разделяют понятия перевода, адаптации и кросс-культурной валидации. Перевод в интерпретации Эпштейна, Рут Миюки Сантоб и Гуллемина определяется как «единий процесс создания анкеты из исходной версии на целевом языке», при этом под адаптацией подразумевается «процесс рассмотрения смысловых и культурных различий между исходной и целевой версиями опросника». Процесс кросс-культурной валидации направлен на проверку сохранения тестом исходных психометрических свойств [115], [81]р.12.

Таким образом, несмотря на существование разных подходов к определению понятия адаптации психометрического теста, все они выделяют схожие элементы процесса адаптации: проверка существования конструкта, для измерения которого разработан тест, в культуре, к которой он будет адаптироваться; выбор переводчиков и критериев оценки их деятельности; непосредственно перевод; объединение разных версий перевода и проверка эквивалентности исходной версии теста переведенной. Необходимо подчеркнуть, что в нашей работе мы сосредоточились именно на последнем из аспектов процесса адаптации.

3. Ключевыми работами, которые посвящены изучению метода когнитивного интервью для задачи адаптации психометрических тестов, выступает работа американских исследователей Пан и Фонд,

а также работа Виллиса. Пан и Фонд разработали классификацию проблем, которые могут быть выявлены с помощью метода когнитивного интервью в рамках адаптации психометрических тестов: культурные, социальные и лингвистические проблемы. Виллису принадлежит определение когнитивного интервью как одного из инструментов адаптации тестов: когнитивное интервью – это метод, нацеленный изучить то, как целевая аудитория понимает, мысленно обрабатывает и реагирует на материалы, предоставленные исследователем, с особым вниманием на смещения и сбои в этом процессе. Теоретико-методологическим изучением когнитивного интервью также занимались ДиМайо и Ротгеб; Мохорко и Глебец; Джоуб, Туранжо и Смит; Уилсон; Стритт и Смит; Конверс и Прессер; Оксенберг и Калтон; Блэр; Штраус и Корбин; Гербер и Велленс; Бикарт и Фелчер; Канеел и Фовлер; Д.М. Рогозин; А. Ипатова, К. Мануильская.

Когнитивное интервью относится группе методов, именуемых когнитивным анализом. Предмет когнитивного подхода заключается в выявлении “мыслительных процессов, которые активизируются при восприятии вопроса и ответе на него, а назначение – в том, чтобы «установить, что человек думает, отвечая на вопросы интервьюера» [475]. Чаще всего основной целью применения когнитивного подхода является тестирование чернового опросного инструментария.

Когнитивный подход включает в себя две группы методов: когнитивное кодирование и когнитивное интервью. Преимущество когнитивного кодирования заключается в возможности зафиксировать характеристики прохождения интервью и выявить те параметры, которые могут оказывать влияние на его прохождение. Недостатком данной группы методов является дороговизна и высокая степень продолжительности процедуры кодирования и анализа результатов, а также тот факт, что когнитивное кодирование не позволяет проанализировать соответствие интерпретаций вопроса исследователем и респондентом, а также выявить недостатки в формулировках вопросов [Cannel, Fowler, 1996. P.21]. То есть, этот подход не годится для задач адаптации уже созданного теста к новому культурному контексту.

Несмотря на многообразие классификацией методов когнитивного интервью, чаще всего исследователи выделяют две «парадигмы»: «проговаривание мыслей вслух» (*think-aloud*) и «метод вопросов-проб» (*probes*) [43]. Применительно к задаче адаптации психометрического теста парадигма «мышление вслух» является менее затратным инструментом, так как не требует предварительного обучения интервьюеров, подробного ознакомления с содержанием анкеты, а также тщательной проработки гайда интервью. Ограничением «проговаривания мыслей вслух» для задачи адаптации является тот факт, что данный подход не всегда может предоставить информацию об интерпретации содержания вопросов, так как информант не сталкивается с данным вопросом напрямую. Исследователь, перед которым стоит цель адаптации психометрического теста, никогда не может быть точно уверен, получит ли он необходимую информацию или нет. Таким образом, несмотря на снижение вероятности возникновения эффекта интервьюера в парадигме «мышление вслух», мы не можем с уверенностью утверждать, что данный подход будет эффективным для достижения задачи адаптации теста.

Параллельно с мышлением вслух развивалась парадигма «интенсивного интервью» («*intensive interviewing*») или метод «вопросов-проб» (*probes*). Эспосито, Ротгеб, Поливка, Хесс и Кампанелли в качестве основных возможностей применения парадигмы «вопросов-проб» выделяют оценку того, насколько представления респондентов расходятся с определениями, представленными в словарях и

справочниках, выявление неоднозначно воспринимаемых респондентом понятий, а также вопросов, которые не отражают цель исследования, установление вопросов, которые в большей степени отражают цель исследования, выявление того, как различные версии вопросов влияют на ответы респондентов [Esposito, 1993. p. 18-19]. Джоуб, Туранжо и Смит среди преимуществ применения метода когнитивного интервью называют получение более детальной информации о когнитивных процессах, которые происходят во время формирования ответа [190]. Гербер и Велленс считают, что данный метод позволяет оценить сензитивность вопросов по отношению к определенным социальным группам [139]. Таким образом, парадигма «вопросов-проб» может предоставить более полную, исходя из целей исследования информацию, по сравнению с «мышлением вслух», так как данный подход позволяет контролировать процесс прохождения интервью информантом.

Американские исследователи из Бюро переписи населения США Пан и Фонд, опираясь на мысль о неотделимости языка, культуры и общества, разработали следующую классификацию проблем, которые могут быть выявлены с помощью когнитивного интервью в рамках адаптации инструментария: лингвистические, культурные и социальные проблемы. Лингвистические проблемы относятся к затруднениям информантов, вызванными использованием устаревших в языке понятий, грамматическими ошибками, особенностью порядка слов или структуры вопроса в целом. Культурные проблемы возникают из-за разных способов выражения одного и того же понятия в культурах. Социальные проблемы связаны с существующими в обществе практиками и социальными институтами [295]. Таким образом, разработанная Пан и Фонд классификация соотносится к тремя видами эквивалентности, для достижения которой необходимо привлечение носителей культурных практик, а также с подходом социолингвистики.

Основным ограничением работы Пан и Фонд является тот факт, что в своем исследовании они используют существующие проблы, не нацеленные на выявление описанных ими проблем. Таким образом, для преодоления данного ограничения когнитивного интервью применительно к задаче адаптации психометрического теста, мы предполагаем, что целесообразным будет развитие данного метода.

#### **4.5 Методы сбора и обработки данных**

В качестве подхода была выбрана качественная методология, а именно когнитивный анализ. Основной целью нашей работы является выявление ограничений и возможностей развития методики когнитивного интервью для задачи культурной адаптации психометрического теста как инструмента измерения компетенций. Основное назначение культурной адаптации психометрических тестов заключается в выявлении смысловых и культурных различий между его исходной и адаптированной версиями, в то время как предмет когнитивного подхода заключается в анализе семантического соответствия вопросов изучаемому предмету. Таким образом, методология когнитивного анализа может выступить в качестве инструмента выявления смысловых и культурных особенностей восприятия информантами вопросов психометрического теста, что является необходимым условием для полноценной адаптации инструментов [474]p.18. В этом отношении цель нашей работы – развитие методики КИ.

В качестве метода исследования было выбрано когнитивное интервью в «парадигме» «вопросов проб», так как именно данный метод (по сравнению с «мышлением вслух») позволяет целенаправленно выявлять смысловые и культурные особенности восприятия вопросов информантами, а также

контролировать процесс интервью.

В качестве определения понятия «когнитивное интервью», мы будем использовать подход Виллиса: когнитивное интервью – это метод, нацеленный изучить то, как целевая аудитория понимает, мысленно обрабатывает и реагирует на материалы, предоставленные исследователем, с особым вниманием на смещения и сбои в этом процессе [Willis, 2005]. Данный подход к определению когнитивного интервью согласуется с понятием адаптации в нашем исследовании как процесса, нацеленного на получение информации о смысловой интерпретации вопросов информантами.

Под развитием когнитивного интервью мы будем понимать использование данной методики для задачи новой проблематизации адаптации психометрического теста, а не только апробации чернового инструментария; разработку зондов, релевантных для задачи культурной адаптации психометрического теста; выявление возможностей и ограничений модифицированной методики когнитивного интервью для задачи культурной адаптации психометрического теста.

**Дизайн исследования.** За основу работы был взят психометрический текст Р\*А, используемый для оценки профессиональных компетенций сотрудников при приеме на работу. В зарубежной практике этот тест используется на протяжении последних 20 лет. Также эта методика удовлетворяла нашим требованиям. При выборе инструмента для его дальнейшей адаптации мы руководствовались такими принципами, как надежность и валидность инструмента, тест не должен был ранее переводиться на русский язык и адаптироваться к российскому контексту, разработчики инструмента согласны предоставить содержание психометрического теста. От компании, которая занималась переводом теста и методических рекомендаций к нему, мы получили ограниченное количество вопросов для когнитивного тестиирования (21 шт.), в отношении формулировок которых эксперты не могли прийти к соглашению. Условием передачи вопросов компанией являлось подписание договора о неразглашении коммерческой тайны.

В качестве основного метода сбора данных выступили полуструктурированные когнитивные интервью, проведенные в рамках парадигмы вопросов-проб.

Полевой этап реализовывался в два этапа. На первом нами было проведено 20 интервью, 10 интервью с “программистами” и 10 интервью с “филологами”. Каждая целевая группа была разделена на две подгруппы, каждой подгруппе был предоставлен свой вариант анкеты с разными альтернативами перевода вопросов. В качестве генеральной совокупности в нашем исследовании выступали студенты 3 и 4 курса бакалавриата, а также 1 и 2 курса магистратуры НИУ “ВШЭ” в г. Москва, так как на данном уровне обучение (незаконченное высшее или высшее образование) целесообразно проводить измерение компетенций. Нами рассматривались только информанты, обладающие опытом работы, так как анкета подразумевает прохождение информантами с опытом работы. В рамках выборочной совокупности нами были выделены 2 контрастные целевые группы: “программисты” (студенты, обучающиеся по направлениям подготовки «Прикладная математика и информатика»; «Информатика и вычислительная техника»; «Информатика и вычислительная техника»; «Программная инженерия»; «Информационная безопасность»; «Информационная безопасность») и “филологи” (студенты, обучающиеся по направлениям подготовки «Филология»).

После окончания этого этапа мы проанализировали адекватность подхода, корректность работы зондов и всей методологии исследования в целом. После внесения корректировок в подход, были проведены 40 интервью. Из них:

- Интервью в формате видеоконференций (20 интервью, в том числе 10 интервью с респондентами моложе 25 лет и 10 интервью с респондентами старше 50 лет; дополнительно варьировался уровень образования – высшее / среднее),
- Интервью в формате «веб-пробирования» (20 интервью, в том числе 10 интервью с респондентами моложе 25 лет и 10 интервью с респондентами старше 50 лет; дополнительно варьировался уровень образования – высшее / среднее).

Для выявления “проблемных” вопросов мы использовали несколько типов «зондов»:

1. “Уверенность в ответе”, когда информант оценивал свой ответ по шкале от 1 до 5, где 1 - абсолютно уверен в ответе и 5 - абсолютно не уверен в ответе. Данный зонд использовался для индикации наиболее проблемных формулировок. Интервьюер задавал дополнительные зонды к вопросам, в ответе на которые информант был не уверен.
2. Метод парадигма (информант самостоятельно пытается переформулировать вопрос при помощи собственных лексических конструкций).
3. Уточняющий зонд (информанта просят объяснить причины выбора своего ответа).

На основе общего анализа ответов информантов после применения метода парадигма и уточняющего зонда мы оценивали семантическую адекватность для каждого вопроса по шкале от 1 до 3: 1 - адекватный ответ (уверенный ответ, без искажения смысла заданного вопроса), 2 - респондент сомневается (есть некоторые отклонения ответа от конструкции вопроса, респондент предлагает свои варианты ответа), 3 - неадекватный ответ (респондент отказывается делать выбор из предложенных вариантов, искажает смысл вопроса, дает свой ответ) [475].

3. «Трехчастный зонд». Для некоторых, заранее отобранных вопросов, мы использовали “смыслоориентированный зонд», состоящий из трех частей, чтобы выявить несоответствие между опытом респондента и его определением обсуждаемого в вопросе концепта. Мы задавали уточняющие вопросы о том, как понимается концепт («светский разговор»), есть ли у него личный опыт и в каком контексте происходило подобное действие.

Кроме того, в конце интервью мы использовали еще два зонда:

- «сложные вопросы». У информанта спрашивали, какие вопросы ему показались сложными, на какие вопросы трудно ответить.
- «зонд на сензитивность». У информанта спрашивали, о каких вопросах ему неловко говорить, какие вопросы было бы некомфортно обсуждать с незнакомым человеком.

## 4.6 Обсуждение

Возможности и ограничения применения «трехчастного зонда». Основные возможности применения «трехчастного зонда» для задачи адаптации психометрического теста как инструмента измерения компетенций заключаются в выявлении смысловых интерпретаций вопросов и смещений в этих интерпретациях с помощью соотнесения личного опыта информанта и его определений ключевых

концептов в обсуждаемых вопросах. Сбор данной информации является необходимым условием для принятия решения о достижении эквивалентности между исходной и адаптированной версиями тестов, что может повысить качество адаптации психометрических тестов. Результаты проведенного нами исследования продемонстрировали тот факт, что «трехчастный зонд» может выявлять потенциальные проблемы в вопросах именно с точки зрения процесса адаптации. Несмотря на то, что при разработке «трехчастного зонда» предполагалось, что данный инструмент будет направлен преимущественно на выявление социальных проблем, результаты анализа интервью продемонстрировали, что «трехчастный зонд» способен также выявлять и лингвистические затруднения участников.

В качестве основных ограничений применения «трехчастного зонда» необходимо выделить невозможность выявления культурных проблем о разных способах выражения одного и того же понятия в исходной и целевой культуре без наличия результатов когнитивных интервью или смысловых интерпретаций вопросов в исходной культуре.

Применение «трехчастного зонда» к большому количеству вопросов затруднительно из-за объемной конструкции данного инструмента, а также из-за факта обучения информантов: в нашем исследовании к концу каждого интервью информанты отвечали на все три части зонда, не дожидаясь вопросов интервьюера.

Специфика исследуемого нами психометрического теста заключается в краткости формулировок вопросов, из которых он состоит. Этот факт облегчает выделение ключевых понятий в вопросах для дальнейшего применения «трехчастного зонда» и составления ожидаемых проб «трехчастного зонда» перед проведением интервью. Применительно к опросному инструментарию, в рамках которого встречаются громоздкие конструкции вопросов, в которых могут быть несколько ключевых понятий, процедура подготовки к интервью осложняется составлением условных проб для сохранения стандартизации, либо появляется необходимость использования эмерджентных проб в зависимости от течения разговора интервьюера и участника, что понизит стандартизированность процедуры проведения интервью.

Использование когнитивного интервью как одного из методов адаптации и «трехчастного зонда» как основного инструмента когнитивного интервью повысит временные и денежные затраты для адаптации теста.

Основными ограничениями нашей работы является апробация развитой методики только на двух контрастных социально-демографических группах, отсутствие возможности сравнения полученных смысловых интерпретаций со смысловыми интерпретациями вопросов носителей культурных практик в англоязычной культуре для апробации «трехчастного зонда» как инструмента выявления культурных проблем, а также субъективность при выделении потенциальных проблем в вопросах теста. Использование зонда «понятность вопросов» в качестве индикатора для применения «трехчастного зонда» к вопросам, которые были малопонятны информантам, требовало введения уточняющих вопросов о причинах низких оценок «понятности» на этапе применения данного зонда.

## 4.7 Результаты исследования

В результате проведения исследования была выявлена эффективность разработанного нами «трехчастного зонда» по трем критериям: выявление информации, соответствующей функциям зонда, количество и

качество обнаруженных проблем.

Интервью в формате видеоконференций (средний показатель 2,8) более эффективные с точки зрения предоставления информации о смысловых интерпретациях вопросов информантами и личном опыте участников взаимодействия с обсуждаемым в вопросе конструктом, по сравнению с «веб-пробированием» (средний показатель – 2,4).

Интервью в формате видеоконференций также более эффективны, по сравнению с «веб-пробированием» с точки зрения выявленных проблем качественно и количественно (в видеоконференциях было выявлено 58 проблем, из них 50% социальных и 50% лингвистических, при «веб-пробировании» - 19 проблем, из них 47% социальных и 53% лингвистических).

Таким образом, наше предположение о меньшей эффективности «веб-пробирования» по сравнению с проведением интервью в формате видеоконференций на базе платформы Zoom подтвердилось.

Исходя из количества и качества выявленных проблем, наше предположение о преимущественном выявление «трехчастным зондом» социальным проблем подтвердилось.

В разрезе наличия опыта участия у информантов в социологических и маркетинговых исследованиях, значительных отличий в показателях выявления необходимой информации «трехчастным зондом» нет. С точки зрения выявления проблем количественно и качественно, и в видеоконференциях и при проведении «веб-пробирования» интервью с участниками без наличия соответствующего опыта выявило больше лингвистических и социальных проблем (в видеоконференциях: 32 проблемы выявлены в интервью с участниками без опыта, 26 – с опытом; при «веб-пробировании»: 12 проблем выявлено в интервью с участниками без опыта, 7 – с опытом).

Таким образом, наше предположение о большей эффективности «трехчастного зонда» при проведении интервью с информантами, у которых есть опыт участия в социологических исследованиях и присутствовали определенные ожидания по поводу интервью, не подтвердилось. Специфика взаимосвязи наличия опыта участия в социологических и маркетинговых исследованиях и эффективности «трехчастного зонда» требует проведения дополнительных исследований.

«Веб-пробирование» позволило избежать возникновение эффекта интервьюера и снизило временные и денежные затраты, необходимые для проведения интервью и составления транскриптов.

Один из выявленных нами недостатков «веб-пробирования» состоит в ограничении применения зонда «понятность вопроса» из-за невозможности использования дополнительных вопросов для уточнения причин выставления низких оценок по данной пробе информантами. Формат «веб-пробирования» также ограничивает использование зонда «понятность вопроса» в качестве индикатора для дальнейшего применения «трехчастного зонда» к вопросам, которые были малопонятны участникам. Невозможность использования дополнительных и уточняющих вопросов относится также в целом ко всей процедуре «веб-пробирования», что снижает информативность полученных ответов, по сравнению с форматом видеоконференций.

Среднее время, затраченное участниками на заполнение анкеты при участии в «веб-пробировании» больше, чем для видеоконференций. Участники исследования в качестве недостатков процедуры «веб-пробирование» отметили длительность процедуры, однообразие вопросов и трудность набора большого количества текста на клавиатуре.

Несмотря на вероятность возникновения эффекта интервьюера, применение «трехчастного

зонда» в формате видеоконференций открыло перед нами возможность задавать уточняющие вопросы информантам, а также контролировать прохождение интервью и анализировать невербальные реакции участников. Некоторые информанты, принявшие участие в «веб-пробировании» в качестве комментариев, отметили тот факт, что им была не понятна цель исследования, в то время как при проведении видеоконференций такой проблемы не возникало. Возможно, данный факт связан с тем, что при прохождении «веб-пробирования» информанты могли не читать приветственное слово, в то время как в видеоконференциях они были вынуждены выслушать вступление информанта с объяснением целей исследования «принудительно».

Ограничением использования видеоконференций в нашем исследовании выступили плохое качество Интернет-соединения и динамиков у некоторых информантов, в связи с чем проанализировать некоторые проведенные интервью не удалось.

Основные направления будущих исследований могут быть связаны с кросс-культурным использованием «трехчастного зонда» в качестве инструмента выявления не только социальных и лингвистических, но и культурных проблем в рамках задачи адаптации опросного инструмента; проведение опросного эксперимента для выявления влияния выявленных нами проблем на психометрические свойства теста: валидность и надежность; проведение исследования на большей выборке испытуемых для изучения особенностей использования «трехчастного зонда» в разных социально-демографических группах; измерение когнитивной усталости информантов, к которым применяется «трехчастный зонд»; уточнение требований администрирования процесса применения «трехчастного зонда» с точки зрения поведения интервьюера, апробация использования «трехчастного зонда» в других форматах проведения интервью, а не только в формате видеоконференций и «веб-пробирования». Отдельным направлением стоит выделить задачу использовать методы сетевого анализа текстовой информации, чтобы развивать аналитические возможности работы с материалами, полученным с помощью когнитивного интервью.

#### **4.8 2.5 Стратегия качественного сетевого анализа: основные подходы, возможности и потенциал применения в прикладных исследованиях**

Современные социологические проблемы связаны с разнообразием методологических исследовательских практик, их обозначением в виде отдельных методологических подходов и их смешиванием с другими методами в эмпирических исследованиях. Одним из таких подходов является сетевой анализ в социологии, который обозначается как анализ структур, состоящих из единиц и связей между ними. Сетевой анализ относят к статистическому методу, поскольку он направлен на изучение социальных структур, где акторы и их отношения встроены в сети, а структура сети оказывает влияние на социальные взаимодействия акторов. Существует практика использования качественных методов в сетевом анализе, где на этапе сбора и анализа данных уделяется внимание именно восприятию и конструированию отношений между акторами в сети [460]. Глубинные социальные структуры, как предмет исследования в сетевом анализе, могут быть обозначены и в качественной методологии, где совмещается структуралистская модель с интерпретативной моделью социологического объяснения [454]. Качественный метод способен отразить конструирование форм отношений и восприятий их акторами.

На данный момент позиция качественной сетевой методологии не определена, поскольку исследователи имеют разные точки зрения на этот счет. По мнению одних авторов, качественный сетевой

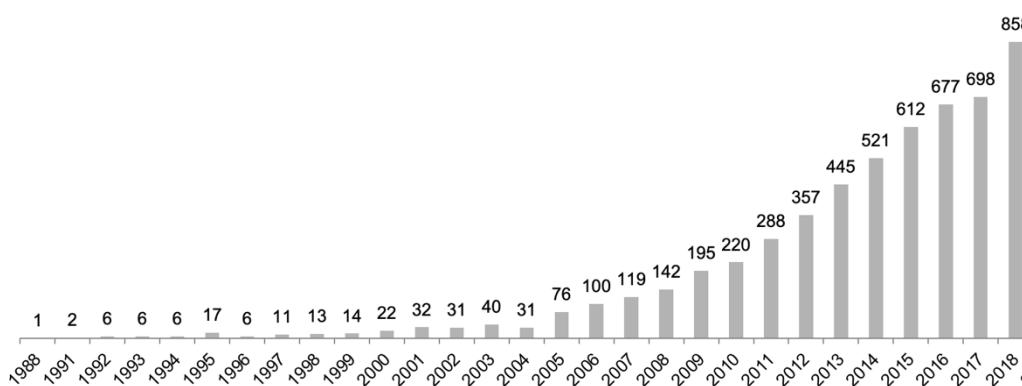
анализ (далее КСА) возможен как «качественное исследование новых типов сетей и стратегий сети, к которым затем можно подключить формальный сетевой анализ», то есть качественный сетевой анализ выполняется поэтапно: интерпретативная часть анализа, а затем ее структурная визуализация [5]. Таким образом, КСА можно представить как двухэтапную методологию, совмещающую интерпретативные возможности со структурацией. С другой стороны, существует мнение об отказе КСА права на методологическую самостоятельность. Такие авторы как Р. Диаз-Боун видят возможность качественной интерпретации сети не более чем в рамках смешанных методов исследования [102]. По его мнению, интерпретативная составляющая качественного сетевого анализа стоит на основе количественного сетевого анализа.

Таким образом, на мой взгляд, границы, проходящие между качественной сетевой методологией и смешанной методологией исследования - весьма размыты, что приводит к неоднозначному обозначению методологических подходов, а также к разноплановым исследовательским практикам использования качественных методов в сетевых исследованиях. Проблема определения и различия качественного сетевого исследования является актуальной, а именно, каким образом сетевое структурирование в качественной методологии может быть обосновано и развернуто как самостоятельная методология.

#### 4.8.1 Становление КСА в социальных науках: литературный обзор

При подготовке литературного обзора в целях систематизации источников литературы по КСА был использован систематический обзор литературы [460] и библиометрический метод [201] сбора и анализа литературы, в рамках которого применялся алгоритм поиска основных путей (Search Path Count) и анализировались сети со-встречаемости ключевых слов. На основе систематического и алгоритмизированного подходов к обзору литературы, была описана теоретическая база КСА.

Количество работ, посвященных КСА в WoS растет начиная с 1990-х и значительно увеличивается с 2000-х (рис. 1). В 2018 году число работ равно 859 и в 2020 году увеличивается до 937 в год. Такой рост опубликованных работ отражает растущий интерес к данной методологии и может обозначать



формирование поле КСА в целом.

Рис. 1. Распределение числа опубликованных работ по КСА по годам

В библиометрическом исследовании поля КСА, удалось выявить основных авторов, ключевые работы, развитие поля в динамике и динамический анализ сетей со-встречаемости ключевых слов. Данные состоят из статей из WoS на основе поиска по ключевым словам - \*“Social network\* + (Qualitative OR Mixed method)” (всего 21,823 публикаций).

На основе алгоритма поиска основных путей (Search Path Count) выявлено два основных научных

поля развития КСА – в социальных и эпидемиологии (Рис. 2 и 3). На рисунке 2 КСА в социальных науках начинается с развития теоретико-методологической фазы с работами М. Грановеттера о экономическом действии, опубликованной в 1985 году и А. Портеса о социальном капитале. Далее следует эмпирическая фаза, где основными объектами выступают мигрантские и преподавательские сети начиная с 2010 года. Преподавательские сети меняют фокус на неформальное обучение в 2015 году, а в объект мигрантских сетей добавляются транснациональные особенности сети. Что касается развития КСА в эпидемиологических науках, то также начинается с теоретико-методологической фазы в 1993 году с работы Морриса о возможности изучения эпидемиологии при помощи сетевого анализа (Рис. 3). Далее с 1999 начинается эмпирическая фаза изучения сексуального контакта, но затем в 2005 году обратно возвращается теоретико-методологическая фаза изучения эпидемиологических исследований при помощи сетевого анализа. В 2011 году снова начинается эмпирическая фаза развития КСА, где объект меняется с сексуального контакта на социальный контакт.

## QSNA in social science

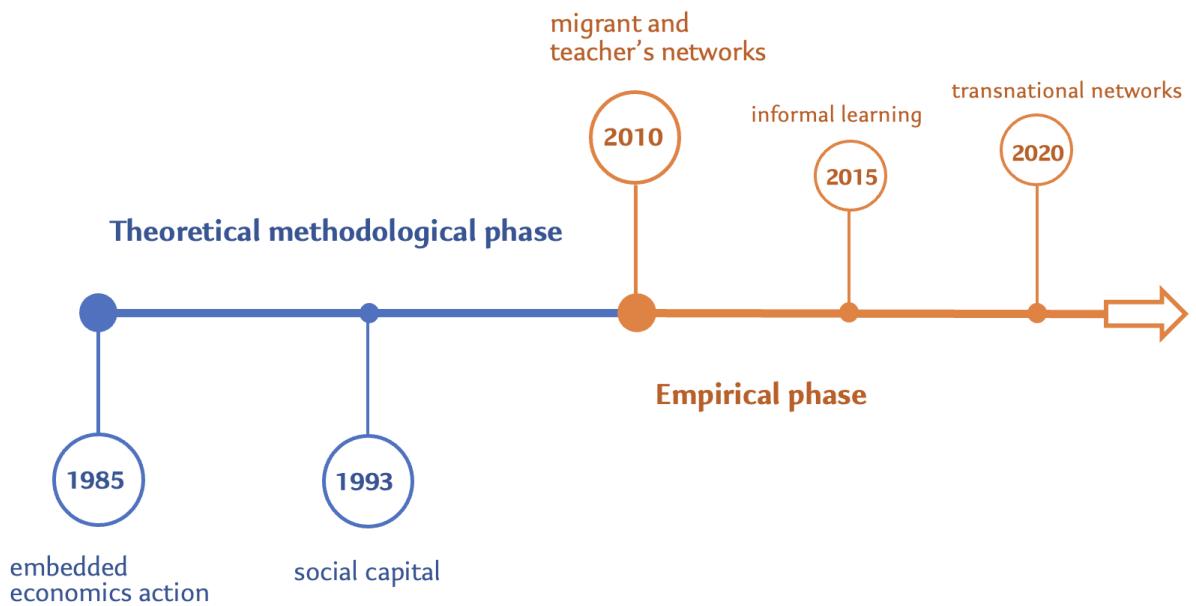


Рис. 2. Развитие КСА в социальных науках

## QSNA in medical science

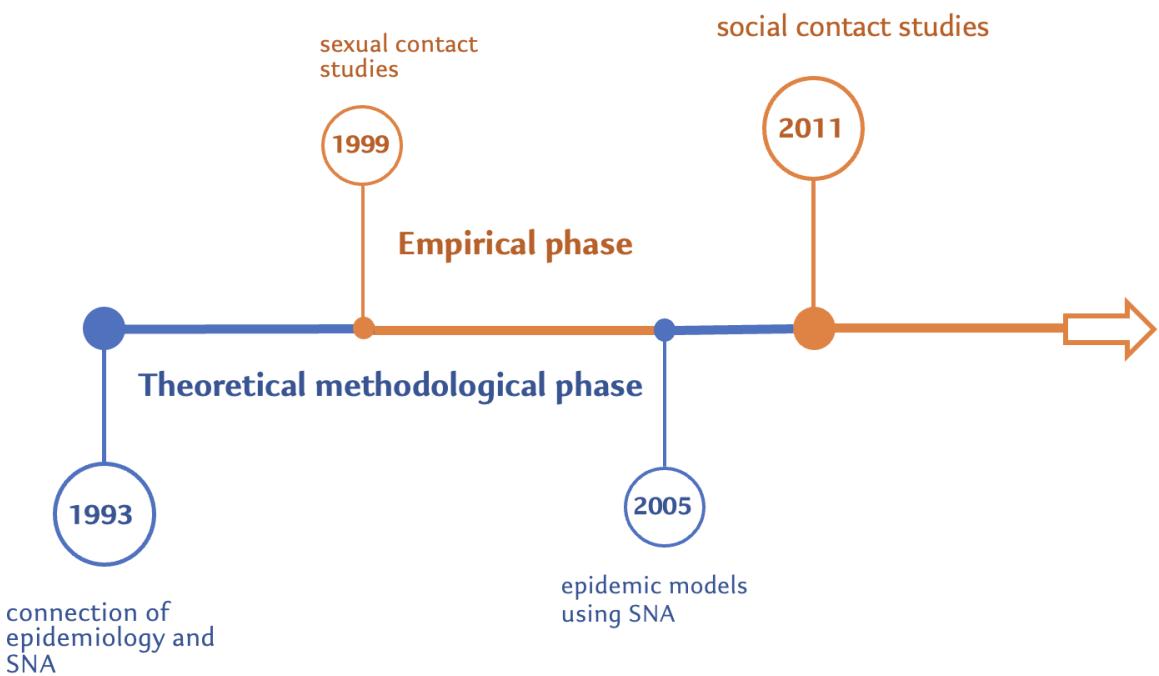


Рис. 3. Развитие КСА в медицинских науках

Показательной может быть сеть со-встречаемости ключевых слов в поле КСА, изображенной на рисунке 4. Сеть со-встречаемости ключевых слов состоит из 20,210 узлов, где связь между узлами обозначает встречаемость в одной и той же статье. Мы выделили только подгруппу из 50 ключевых слов, которая состоит из самых важных узлов, которые ближе всего друг к другу. Центральным ключевым словом является *social*, который связан с несколькими тематиками в сетевом анализе, интернет, медиа коммуникации, социальный капитал, образование и поддержка. Ключевое слово *qualitative* связано с методологическими темами, такими как *research* и *study* и эмпирическими объектами – *health*, *mental*, *care*.

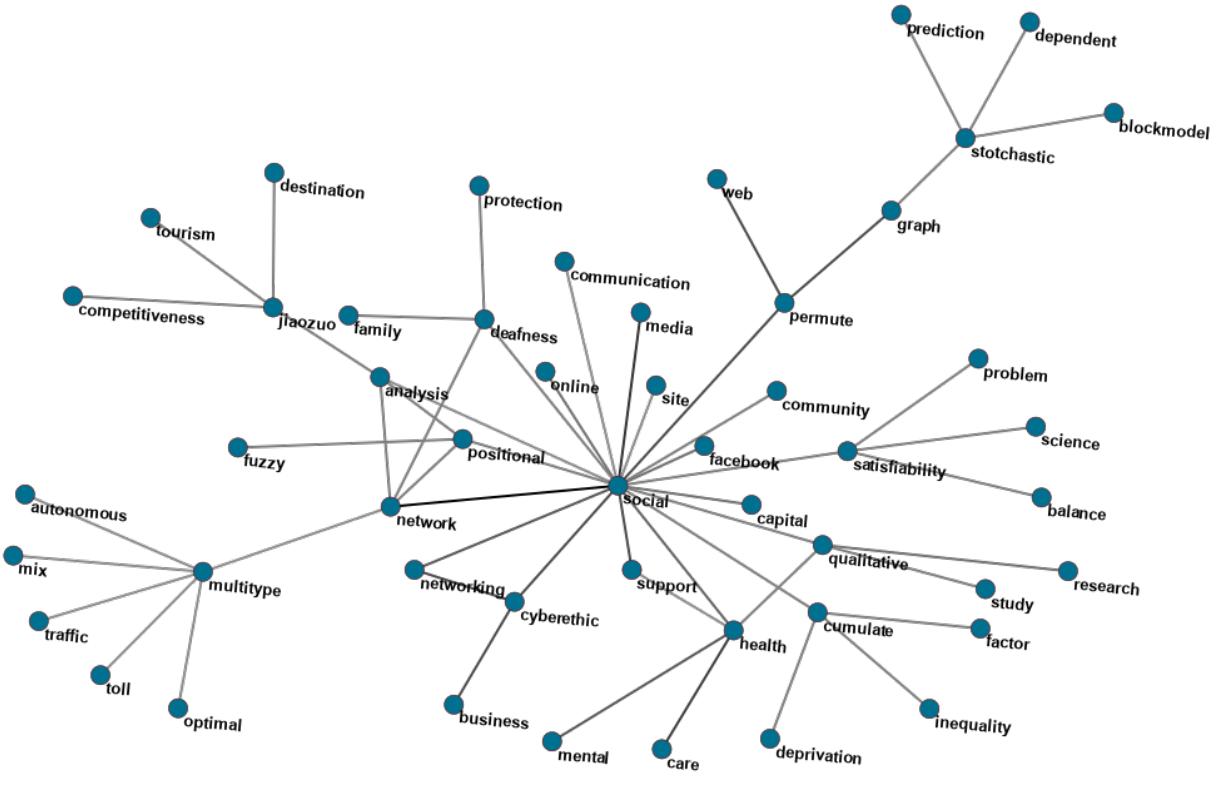


Рис. 4. Сеть со-встречаемости ключевых слов в поле КСА

В целом, приводя сравнение с развитием поля сетевого анализа в работе Фримана [128], КСА проходит такие же этапы: 1) Этап формирования на основе классических социологических работ, 2) «Рождение и смерть» сетевого анализа в поле антропологии и социальной психологии, 3) «Темные века» сетевого анализа, когда он развивался в изолированных друг от друга дисциплинах, 4) «Взросление» и институционализация сетевого анализа - появление конференций, научных журналов, компьютерных программ, образовательных программ в университетах. На данный момент КСА также институционализируется, поскольку качественные сетевые исследования начинают появляться в программах конференций, научных статьях.

#### 4.8.2 Теоретико-методологические основания КСА

КСА относится к сетевым исследованиям, в основе которых лежит использование сетевого подхода. Сетевой подход может быть определен как «комплекс теоретико-методологических направлений, использующих понятие сети для объяснения социальных явлений» [464]. Сетевой подход развивается в двух исследовательских направлениях — формальном сетевом анализе, который известен как анализ социальных сетей (SNA, social network analysis), и реляционистском, или отношенческом, подходе, относящемся к реляционной социологии. Анализ социальных сетей позволяет определять взаимосвязи между единицами анализа, выявлять глубинные социальные структуры и находить закономерности взаимодействия в социальных группах. Реляционистский подход, относящийся к реляционной социологии, восходит к исследованиям культуры, где социальная реальность изучается через призму социальных отношений, разворачивающихся в динамике и неотделимых от контекста. Таким образом, КСА также базируется на структуралистские и реляционные подходы, как и сетевой анализ.

Качественный сетевой подход направлен на выявление глубинных смыслов отношений в сети.

Под глубинными смыслами отношений понимается неявная интерпретация взаимодействий индивидов, находящихся в общем сетевом пространстве. Смыслы отношений в сети как значения социальных сетей в целом изучает Ян Фузе в своей книге «Social networks of meanings» [133], где он развивает теоретическую базу для сетевого анализа, намеренно избегая структуралистских оснований. Смыслы сетевых отношений Фузе связывает с веберовской категорией «понимания», которая занимает интерпретирующую позицию объяснения социального явления. Фузе рассматривает ряд теоретических концепций, имеющих отношение к полю анализа социальных сетей (Табл.1).

Табл. 1. Теоретические концепции к КСА

<b>Теоретический подход</b>	<b>Ключевые авторы</b>	<b>Суть концепции</b>
Прагматический и интеракционистский подходы	Мустафа Эмимбайер, Ник Кроссли, Джон Леви Мартин	- поведение и социальные структуры имеют отношение к субъективным состояниям и процессам- социальные отношения формируются в интенсивном взаимодействии двух субъектов лицом к лицу, что ведет к обмену символами и общности взглядов (Кроссли),- важность культуры и индивидуальной свободы действий в социальных сетях (Эмирбайер)- определение сетевых процессов с помощью позиций в сети (Мартин)
Реляционная социология Харрисона Уайта	Харрисон Уайт	- социальные сети понимаются как смысловые структуры, которые вводятся в действие и согласовываются в процессе коммуникации,- коммуникация происходит в историях об идентичностях, при “транзакциях” между различными социокультурными контекстами, а также при использовании культурных рамок для определения отношений

<b>Теоретический подход</b>	<b>Ключевые авторы</b>	<b>Суть концепции</b>
Реляционная работа	Вивиан Зелизер	-люди договариваются и разграничают свои личные отношения со ссылкой на такие социальные рамки, как дружба, любовь, товарищеские отношения и т.д.
“Другая” реляционная социология	Пьерпаоло Донати,Маргарет Арчер	- социальные отношения представляют собой реальность <i>sui generis</i> , т.е. общество состоит из отношений и социология должна рассматривать отношения как онтологические и эпистемологические отправные точки,-социальные отношения определяются как “эмерджентный эффект” действий множества действующих лиц
Акторно-сетевая теория	Бруно Латур,Мишель Каллон,Аннемари Мол,Джон Лоу	- исследования в области науки и техники, где утверждается, что нужно описывать расположение человеческих субъектов и материальных объектов в терминах ассоциаций, взаимодействий между ними,- оба вида “действующих лиц” (человеческие и материальные) приобретают свое значение только в процессах, происходящих между ними,- научные открытия являются результатом действий, выполняемых этими сетями, а не отдельными субъектами.

Теоретический подход	Ключевые авторы	Суть концепции
Теория систем Никласа Лумана	Никлас Луман	-коммуникация протекает самореферентно, опираясь на предыдущее общение и создавая различные виды социальных систем - от личных встреч через официальные организации (компании, университеты, административные единицы) до крупномасштабных функциональных подсистем общества, таких как политика, экономика, юриспруденция и наука,-самопроизводство (аутопоэзис) и производство ограниченных социальных образований (систем)

Для собственной концептуализации смыслов отношений в сети Фузе концентрируется на реляционной социологии Уайта [420] и теории систем Лумана [250], задавая ключевые категории в качестве символических конструкций к пониманию глубинных смыслов отношений в сети. Первая ключевая категория – это *самисмысли отношения*, понимающиеся как совокупность ожиданий относительно поведения индивидов по отношению друг к другу. Эти ожидания в отношениях управляют ходом коммуникации и связаны с другими формами смыслов отношений: идентичностями, социальными ролями (например, гендер), институционализированными ролями (профессор/студент) и культурными моделями отношений (любовь, дружба, покровительство). Все это предписывает определенные виды отношений в сети, и они стабилизируются, если коммуникация в них соответствует этим социальным категориям, ролям или культурным моделям [133].

Другой важной категорией для концептуализации социальных сетей является понятие *коммуникативных событий*. Социальные отношения существуют как готовые конструкции, поскольку они формируются и воспроизводятся в ходе событий. Такие события называются «транзакциями» в реляционной социологии. Опираясь на теорию Никласа Лумана, Фузе констатирует, что социальные сети определены как паттерны *реляционных ожиданий* (реляционных определений ситуации), которые возникают, стабилизируются и изменяются в ходе *коммуникативных событий*. Процесс коммуникации влечет за собой приписывание событий индивидам, что приводит к конструированию ожиданий относительно их поведения со стороны других. Такие коммуникативно сконструированные ожидания отношений составляют смысловую структуру социальных сетей и обеспечивают наблюдаемые закономерности коммуникации.

Основным процессом всего социального является *коммуникация*, которая определяется как обработка смысла в коммуникативных событиях [250]. *Коммуникативные события* основаны на речевых и неречевых формах взаимодействия. К особенно важным коммуникативным событиям

относят неречевые, например, жесты, мимика, рукопожатия, объятия, поцелуи и т.д. Таким образом, многие взаимодействия обозначаются как коммуникативные события, если они передают информацию и связывают действующих лиц. *Коммуникативные события* оставляют след в социальном мире не благодаря намерениям, знаниям или другому субъективному значению действующих лиц, а благодаря тому, что их понимают и реагируют на них.

Таким образом, Фузе, концентрируясь на интерпретативном понимании социальных сетей, определяет основание смыслов отношений в сети в коммуникационных событиях (или транзакциях), влияющих на процесс коммуникации как основного процесса всего социального (Рис.5). Конструирование смыслов отношений происходит за счет множества транзакций, исходя из чего формируются реляционные ожидания относительно хода коммуникации и возможного прогноза следующих транзакций.



Рис. 5. Аналитическая схема

понимания социальных сетей как глубинных смыслов отношений через коммуникативные события по Фузе

Я предлагаю свою аналитическую схему изучения глубинных смыслов отношений в КСА на основе концептуализации Фузе. На мой взгляд, в аналитическую схему следует добавить быть добавлен категориальный компонент *контекста отношений*, способный обобщать исходную коммуникацию в единую форму. Поскольку в прежнем дизайне не хватает категории, которая будет накапливать опыт коммуникативных событий. Где в свою очередь, рефлексирующие индивиды обозначают смыслы отношений исходя из контекста этих отношений, заданных на основе коммуникативных событий и реляционных ожиданий. Также немаловажным элементом выявления глубинных смыслов отношений является *рефлексивность* актора. *Рефлексивность* понимается как способность индивида осознавать социальные отношения, наделять их эмоциями, личными историями и обобщать в сконструированный паттерн. Можно предложить дополнить аналитическую схему понятием *социальной структуры* в понимании Гидденса в теории структурации, где *социальная структура* поддерживается через действие актора, а действие приобретает смысл только в контексте структуры [143]. *Рефлексирующий актор* своими действиями формирует и воспроизводит *социальную структуру*, а также он способен изменить структуру, поэтому *социальная структура и отношения* в ней рассматриваются в динамике. Отношения между акторами рассматриваются согласно структурирующим принципам Гидденса: сигнификация, легитимация и власть, то есть между акторами всегда формируется иерархия. Аналитическая схема

изучения глубинных смыслов отношений, дополненная вышеупомянутыми терминами, изображена на Рисунке 6.



Рис. 6. Аналитическая схема изучения глубинных смыслов отношений в социальной сети

Таким образом, глубинные смыслы отношений как предмет качественного сетевого анализа обозначены как неявные конструкции понимания сути отношений между акторами в сети. Осознание этих глубинных смыслов возможно в связи с рефлексивностью актора, его анализом предыдущих коммуникативных событий или транзакций и отношений в целом. По Гидденсу социальная структура воспроизводится в действиях рефлексирующего актора, как и сам актор воспроизводит действия согласно социальной структуре. Такое взаимодействие отражено и в теории систем Лумана, где происходит самопроизводство или аутопоэзис систем. Коммуникативные события или транзакции накапливаются в контексте отношений, исходя из которого рефлексирующий актор понимает смыслы отношений в сети. Контексты и коммуникативные события согласуются с социальной структурой, так как актор воспроизводит свои взаимодействия или транзакции согласно принципам, заданными социальной структурой. В последнем слое аналитической схемы обозначена связь коммуникации и отношений, где показано влияние коммуникации между акторами на установку отношений между ними. Однако, погружаясь внутрь формирования отношений, выявляется связь между коммуникативными событиями и социальной структурой через рефлексивность актора.

#### 4.8.3 Место качественного сетевого анализа в сетевых методологических подходах

Данный раздел посвящен поиску места качественного сетевого анализа в методологических подходах в сетевых исследованиях. Структурирование методов в сетевых исследованиях возможно через классическое социологическое разделение на количественные и качественные методы анализа данных, также посредством альтернативного сетевого разделения на количественные методы, качественные

методы и эго-сетевой анализ, предложенное Фузе и Мютцель [134]. В социологии разделение на количественные и качественные методы является одним из основополагающих методологических принципов. Качественные методы предполагают использование статистического категориального аппарата и направлены на поиск взаимосвязей между переменными, выявление причинно-следственных зависимостей и другие способы объективного объяснения социальной реальности на макроуровне. Качественные методы предлагают интерпретативный подход к объяснению социальной реальности и направлены на понимание социальных феноменов на микроуровне. Качественный сетевой анализ (или формальный сетевой анализ) подразумевает структурный подход к изучению социальной реальности через поиск взаимодействий акторов в общей сети. Качественный сетевой анализ направлен на выявление смыслов отношений акторов в сети взаимодействий.

Другой способ разделения методологических подходов в сетевом исследовании предложен в статье Фузе и Мютцель [134], в котором сравниваются количественный, эго-сетевой и качественный подходы в сетевом исследовании. Эго-сетевой подход понимается как исследование сети с точки зрения конкретного актора или эго. В то же время авторы не обозначают эго-сетевой подход и качественный сетевой подход как отдельные методологические подходы. Эго-сетевой анализ представлен как статистический анализ эгоцентричных сетей, а качественный сетевой анализ подан обобщенно как качественные методы.

Исходя из разделения Фузе и Мютцель [134], сетевой анализ как изначально анализ структур является доминирующим в сетевых исследованиях. Помимо формального количественного сетевого анализа, есть подгруппа эго-сетевого анализа, внутри которого находится блок качественных сетевых исследований – что составляет «русскую матрешку» (Рис. 7). На основе теоретических оснований сетевого подхода в социологии можно выделить классические работы в сетевом анализе, которые можно отнести в целом ко всем сетевым исследованиям – это «библия» сетевого анализа Вассермана и Фауст [415], работы Грановеттера [150], Фримана [127], Эмирбайера и Гудвина [113]. К теоретическим основаниям эго-сетевого подхода авторы относят исследования сообществ [412] и исследования социального капитала [80, 149, 311, 312]. По мнению авторов, качественный подход в сетевом исследовании ссылается на классические социологические труды Зиммеля [1], Вебера [418], Элиаса [110] и Мида [263], а также интеракционистский подход к структуре [123], реляционную социологию Уайта [420] и акторно-сетевую теорию [224].



Рис. 3. Методологические подходы в сетевом исследовании по Фузе и Мютцель

Сравнение методологических подходов в сетевом исследовании приведено в Таблице 2. Предметом количественного сетевого анализа являются глубинные структуры в сети [465]. В эго-сетевом анализе изучается позиция актора в сети. К предмету качественного сетевого анализа можно отнести глубинные смыслы отношений в сети и контексты их взаимодействия [460]. На уровне объекта, в рамках количественного подхода в сетевом исследовании изучаются полные сети взаимодействий. Однако изучение эго-сетей или персональных сетей происходит посредством количественных методов. В качественном и эго-сетевом подходах на уровне объекта исследования рассматриваются только персональные сети взаимодействия. Говоря о моделях объяснения, количественный сетевой анализ, стремится объяснить макросоциальные аспекты социальных явлений, структурируя взаимодействия индивидов. Качественный сетевой анализ можно обозначить как синтез структурного и интерпретативного подходов к социологическому объяснению [454], поскольку в нем присутствует структуралистские объяснения социальных явлений исходя из взаимодействий и отношений, а также субъективные интерпретации сетей взаимодействий. Для количественного сетевого подхода уровнем анализа является макроуровень, в котором рассматривается «взгляд сверху» на целостную полную структуру сети. Эго-сетевой анализ изучает сеть на микроуровне, как и в качественном сетевом подходе, однако предлагает сбор и анализ данных в том же виде как и количественный сетевой подход, где сбор данных может осуществляться как количественным, так и качественным способом и осуществляется количественный анализ данных. Тогда как в качественном сетевом подходе используется только качественный способ сбора данных и используется только качественный способ анализа данных.

Табл. 2. Сравнение методологических подходов в сетевом исследовании

Основания для сравнения	Количественный сетевой подход	Качественный сетевой подход	Эго-сетевой анализ
Предмет	Глубинные социальные структуры	Глубинные смыслы отношений	Позиция актора в сети

<b>Основания для сравнения</b>	<b>Количественный сетевой подход</b>	<b>Качественный сетевой подход</b>	<b>Эго-сетевой анализ</b>
<b>Объект</b>	Сеть, эго-сеть или персональная сеть	Эго-сеть или персональная сеть	Эго-сеть или персональная сеть
<b>Теоретическое основание</b>	Связь с реляционной социологией	Связь с теорией социального капитала	Связь с реляционной социологией, с теорией социального капитала, акторно-сетевой теорией
<b>Модель объяснения</b>	Структурализм	Синтез структурализма и интерпретативизма	Структурализм
<b>Уровень анализа</b>	Макроуровень	Микроуровень	Микроуровень
<b>Сбор данных</b>	Количественный, качественный	Качественный	Количественный, качественный
<b>Анализ данных</b>	Количественный	Качественный	Количественный

Углубляясь в предмет исследования, Фузе и Мютцель [134] раскрывают особенности методологических подходов, разделяя глубинные социальные структуры количественного сетевого анализа и структуру эго-сети в эго-сетевом анализе и качественном сетевом анализе. В социологических исследованиях разделение на качественные и количественные методы подразумевают обособленные типы исследований, в каждом из которых свои предметы исследований и теоретические основания. В количественном исследовании изучаются объективные показатели, которые поддаются измерению, тогда как в качественном исследовании рассматриваются неизмеримые субъективные свойства, в которых исследователь выявляет смысл. Применяя данное разделение на сетевые исследования, можно сказать, что они отличаются по объекту исследования – в количественных сетевых исследованиях изучаются разные виды сетей, тогда как в качественном сетевом исследовании рассматриваются только эго-сети или персональные сети. Получается, что социологическое разделение на количественные и качественные методы является более общим, тогда как разделение на количественный, эго-сетевой и качественный сетевой анализ включает особенности разных исследовательских направлений и дизайнов. Поскольку, с точки зрения предмета в сетевых исследованиях в социологии, количественный анализ подразумевает выявление структуры сети, а в качественном сетевом исследовании фокус делается на смыслы отношений и их контексты.

Исходя из анализа существующих методологических подходов в сетевых исследованиях можно предложить свою систематизацию методологических сетевых подходов. На Рисунке 8 изображена схема методологических подходов в сетевых исследованиях, которые направлены на предметы разных уровней: на макроуровне изучаются сетевые структуры, на мезоуровне сетевые исследования направлены на изучение подгрупп в сети, и на микроуровне проводится эго-сетевой анализ. Эго-сетевой анализ в свою очередь возможен в виде количественного сетевого анализа, где фокус исследования ставится на атрибутах актора, а также в виде качественного сетевого анализа с возможностью фокусироваться на глубинных смыслах отношений в сети.

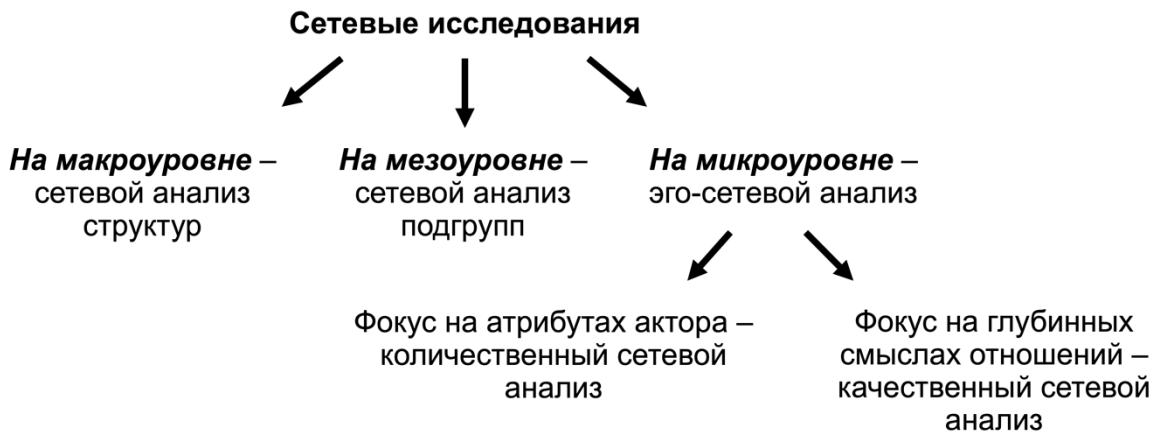


Рис. 8. Методологические подходы в сетевом исследовании

#### 4.8.4 КСА: определение, основные задачи и способы реализации

КСА можно охарактеризовать как методологический подход к изучению отношений в сети. Объектом качественного сетевого анализа являются персональные сети или эго-сети, предметом являются глубинные смыслы отношений в сети. Целью качественного сетевого анализа является выявление глубинных смыслов отношений в сети. Рассмотрев разные способы смешивания методов на этапах сбора и анализа, можно заключить, что когда на этапе сбора и анализа данных применяются именно качественные методы, то в таком случае сетевое исследование является качественным. Для осуществления качественного сетевого анализа, на этапе сбора данных могут применяться те методы, которые могут быть качественно проанализированы - интервью, наблюдение и анализ документов, а также структурированы при помощи построения сетевой карты. Таким образом, качественный сетевой анализ состоит из двух этапов: сбор качественных и сетевых данных, анализ качественных текстовых данных, анализ и визуализация сетевых данных.

Бетина Хольштайн относит качественный сетевой анализ к интерпретативному подходу, который рассматривает социальную реальность как сконструированные, осмыслиенные и отсылающие к контексту действия, зависящую от точки зрения актора и развивающуюся в динамике [175]. Однако на мой взгляд, качественный сетевой анализ можно отнести к синтезу структурализма и интерпретативизма, поскольку в нем есть признаки как интерпретативной версии социальной реальности, так и ее структурный компонент [454]. Объектом качественного сетевого анализа являются персональные сети или эго-сети, предметом являются глубинные смыслы отношений в сети [460]. Глубинность смыслов отношений понимается как суть и восприятие взаимодействий, не лежащих на поверхности, а проявляющихся в глубине. Целью качественного сетевого анализа является выявление глубинных смыслов отношений в сети.

На основе выявленных возможностей качественного сетевого анализа в статье Ким [460] и областей исследования Хольштайн [175] можно выделить основные задачи качественного сетевого анализа:

- Выявлять смыслы отношений в сетях,
- Исследовать и описывать контекст взаимоотношений в сетях,
- Изучать сетевые практики,
- Изучить сетевые ориентации и интерпретации,
- Определять важность общей сети, некоторых кластеров и конкретных узлов,

- Анализировать временные взаимосвязи в сетях и общую динамику сети,
- Фокусироваться на действиях и стратегиях участников для создания сети,
- Получать доступ к труднодоступным сообществам.

На основе различных типов данных и способов их анализа мною было выделено четыре способа реализации КСА в эмпирических исследованиях со смешанными методами (таблица 2). Ввиду изучаемой тематики для каждого способа реализации характерно наличие качественных данных для анализа.

	<i>Тип данных</i>
<i>Качественный</i>	
<b>QUAL → QUAL</b>	+
<b>QUAL → QUAN</b>	+
<b>QUAL → QUAN + QUAL</b>	+
<b>QUAN + QUAL → QUAN + QUAL</b>	+

Табл. 2. Способы реализации КСА в стратегии смещивания методов

**4.8.4.1 QUAL → QUAL** Первый выделенный способ реализации КСА в эмпирических исследованиях со смешанными методами основан на качественных данных, анализируемых качественным образом. Смешивание методов в данном случае подразумевает комбинирование разных качественных методов сбора данных.

При этом способе реализации исследования КСА может рассматриваться как отдельная самостоятельная методология. В исследовании социального капитала мигрантов авторы при помощи сетевой карты в ходе полуструктурированных интервью выделяют основных акторов, которые повлияли на формирование бизнеса в начале и спустя 10 лет [370]. По мнению авторов, данная методика относится исключительно к количественному сетевому анализу, однако на мой взгляд, она может быть применена как к количественному, так и в качественном подходе.

**4.8.4.2 QUAL → QUAN** Второй выделенный способ реализации КСА в эмпирических исследованиях со смешанными методами также основан на качественных данных, однако они конвертируются для количественного сетевого анализа. В исследованиях применяется техника построения сетевой карты на основе качественных данных из интервью определялась как компонент количественного сетевого анализа, поскольку анализ сетевой карты выполнялся с помощью количественного анализа.

**4.8.4.3 QUAL → QUAN + QUAL** Третий выделенный способ реализации КСА в смешанном исследовании характеризуется тем, что качественные сетевые данные анализируются конвертированным способом. Конвертированный тип дизайна смещивания методов проходит три основные стадии: 1) сбор данных осуществляется при помощи интервью или с использованием имеющихся текстовых данных, 2) конвертация сетевых данных для возможности количественного сетевого анализа и 3) анализ данных и интерпретация, где «качественная» информация комбинируется с количественным сетевым анализом на каждой стадии.

**4.8.4.4 QUAN + QUAL → QUAN + QUAL** Четвертый выделенный способ реализации КСА характеризуется тем, что помимо качественных данных предполагает наличие данных количественного

типа, и все они анализируются как качественно, так и количественно. Во всех статьях методом сбора качественных данных является интервью, тогда как количественные методы сбора различаются: часто это опросные данные, но в некоторых исследованиях это собранные данные с сайта и библиографические данные.

Таким образом, из четырех выделенных способов реализации КСА в эмпирических исследованиях два способа (первый QUAL → QUAL и второй QUAL → QUAN) не могут быть отнесены к смешиванию методов, т.к. в них не соблюдаются условия, описанные Домингуес и Гольштейн. При этом первый способ можно отнести к использованию КСА как самостоятельной методологии. Два других способа реализации КСА на практике (третий QUAL → QUAN + QUAL и четвертый QUAN + QUAL → QUAN + QUAL) относятся к смешанным методам. Третий способ можно отнести к конвертированному дизайну смешивания методов – где качественные данные переводятся в количественную форму, чтобы проводить оба типа анализов данных. В четвертом способе были выявлены последовательный и параллельный типы дизайнов смешивания методов.

Немаловажным компонентом применения качественного сетевого анализа является структурализация отношений между акторами при помощи построения сетевой карты. Это техника структурирования социальных отношений, которая может быть выполнена самостоятельно исследователем, либо предлагается информанту во время интервью. Для заполнения сетевой карты во время интервью информанту выдают лист бумаги с пустыми концентрированными кругами, куда их просят вписать имена или инициалы важных персон. За генератором имен следует набор стандартизованных интерпретаторов имен, которые задавали вопросы об атрибутах изменений в круге и типе связи с каждым изменением.

Сбор сетевых карт возможен двумя способами: самостоятельное заполнение информантом или сетевая карта строится самим исследователем, где он полагается на данные из интервью, наблюдения или документов. Эти подходы к сбору сетевых карт сравнивались между собой для выбора наиболее подходящего способа.

Построение своей сетевой карты информантом – это рефлексивное и вдумчивое упражнение. Поскольку интервью проводится онлайн, то нужно выделить время на построение сетевой карты и после на обсуждение этой карты. Безусловно, сам процесс построения сетевой карты интересный и затягивает самого информанта, однако не всегда хватает времени на полноценное обсуждение. Также тратится время на постановку задачи построения сетевой карты и может быть сложно донести образ конечного результата. Помимо этого, есть предположение что сбор сетевой карты онлайн и по зуму может быть сложнее, чем вживую. Как итог, сетевую карту информант отправляет исследователю и качество картинки может быть не таким высоким или на карте присутствуют помарки, подчерк может быть непонятным и др. Однако, построение сетевой карты информантом как упражнение может быть полезным, т. к. интервью может не охватить каких-то людей или не обозначить изменения в отношениях с другими узлами. Получается, что благодаря самостоятельному построению сетевой карты информантом сглаживаются ограничения сбора данных посредством интервью.

Другим способом построения сетевой карты является заполнение сетевой карты самим исследователем после интервью. Этот способ сбора сетевых данных является более экономичным, не тратится время информанта, а также исследователь визуализирует сетевую карту как нужно для его исследования. Для построения сетевой карты необходимо лишь задать дополнительные вопросы

о социальном окружении информанта. Визуализация полностью зависит от исследователя, поэтому, вероятно, она будет высокого качества и наглядно структурировать все контакты. Однако в данном случае, исследователь в большей степени полагается на интервью, в ходе которого информант может вспомнить не всех узлов или отметить изменения не во всех отношениях.

#### **4.9 3.1.1 Программа «Bib-eLib» для сбора и обработки библиографических данных на русском языке из электронной библиотеки eLibrary**

##### **4.9.1 Введение**

В рамках реализации научного проекта лаборатории научным коллективом МЛ прикладного сетевого анализа разработан служебный результат интеллектуальной деятельности – программа для ЭВМ «Программа «Bib-eLib» для сбора и обработки библиографических данных на русском языке из электронной библиотеки eLibrary». Программа «Bib-eLib» разработана Д. В. Мальцевой, В. А. Ващенко, Л. В. Капустиной, А. В. Ким, Т. Е. Щегловой, Л.Г. Ципесом в НИУ Высшая школа экономики, являющейся ее правообладателем (свидетельство о государственной регистрации программы для ЭВМ № 2023684182, регистрация в реестре программ для ЭВМ 14.11.2023) (Рис.1).

**Рис. 1. Свидетельство о регистрации программы «Bib-eLib» для сбора и обработки библиографических данных на русском языке из электронной библиотеки eLibrary**

РОССИЙСКАЯ ФЕДЕРАЦИЯ



## СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2023684182

Программа «Bib-eLib» для сбора и обработки  
библиографических данных на русском языке из  
электронной библиотеки eLibrary

Правообладатель: Федеральное государственное автономное  
образовательное учреждение высшего образования  
"Национальный исследовательский университет "Высшая  
школа экономики" (RU)

Авторы: Мальцева Дарья Васильевна (RU), Ващенко Василиса  
Андреевна (RU), Капустина Лика Владимировна (RU), Ким Арюна  
Витальевна (RU), Щеглова Тамара Евгеньевна (RU), Чипес Лев  
Григорьевич (RU)

Заявка № 2023683161

Дата поступления 02 ноября 2023 г.

Дата государственной регистрации

в Реестре программ для ЭВМ 14 ноября 2023 г.

Руководитель Федеральной службы  
по интеллектуальной собственности

Ю. С. Зубов

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ  
Сертификат 429fb6fe38d3164bc19e6983b730faa7  
Владимир Зубов Юрий Сергеевич  
действителен с 18.08.2023 по 02.08.2024

Цель работы программы «Bib-eLib» – сбор и обработка библиографических данных на русском языке из электронной библиотеки eLibrary для применения в области научометрических и библиометрических исследований, в особенности в области сетевого анализа библиографических сетей.

### 4.9.2 Описание работы программы

Программа «Bib-eLib» характеризуется следующими функциональными характеристиками:

- Программа осуществляет выгрузку массива данных о научных публикациях через API электронной библиотеки eLibrary при наличии предварительно заключенного контракта с Национальной электронной библиотекой НЭБ;
- Программа проводит предварительную обработку массива данных о научных публикациях, выгруженных через API электронной библиотеки eLibrary, в части имен авторов и их аффилиаций (организаций);

- Программа реализовывает алгоритм дизамбигуации авторов публикаций в предварительно обработанном массиве данных путем создания новых универсальных идентификаторов авторов;
- Программа проводит анализ итогового массива данных по основным статистическим метрикам в области библиометрических исследований и визуализировать распределение подсчитанных метрик;
- Программа создает двумодальную сеть связей между научными публикациями и их авторами в формате с расширением .net, пригодном для дальнейшей обработки в программе для анализа и визуализации больших сетей Pajek.

Программа подходит для реализации на Windows, MacOS, Linux на базе интегрированных с ОС графических пользовательских интерфейсов (GUI). Программа написана на языке программирования Python и включает архив файлов формата Jupyter Notebook (.ipynb) размером 266 КБ. Описание файлов .ipynb-files приводится ниже. В качестве примера для сбора взят массив данных о публикациях российских социологов, который был собран в рамках проекта “Паттерны коллаборации в Российском социологическом сообществе” (грант РНФ, руководитель Д.В. Мальцева, 2021-2023 гг.).

1. Файл “0. Выгрузка данных о статьях авторов.ipynb” - в этом файле производится выгрузка данных о статьях конкретных авторов на eLibrary через сервис API Elibrary 011 (не используется напрямую в программе, но может быть полезным). В качестве примера, в этом файле Jupyter Notebook производится выгрузка данных eLibrary по публикациям российских авторов в области социологии. Осуществляется импорт необходимых библиотек, загружаются id авторов, данные о которых необходимо выгрузить, производится выгрузка кратких данных обо всех статьях этих авторов из API Elibrary. Основной код выгрузки к API-011 Elibrary был подготовлен Ликой Капустиной в сентябре-октябре 2022 года. Цель выгрузки - выгрузить краткие справочные данные о статьях всех уникальных авторов, встречающихся в основном массиве данных (но не все данные по статьям). Собранные данные не использовались для дообновления финальных данных проекта, но код в этом ноутбуке может быть полезен для других исследователей. Результат выгрузки – файл all\_final\_authors\_with\_papers.csv, содержащий в себе информацию о всех статьях всех авторов, присутствовавших в 8 колонках наших изначальных данных.
2. Файл “1. Выгрузка данных по статьям.ipynb” – представленный в этом файле код используется для выгрузки данных о статьях через API Elibrary. В этом файле Jupyter Notebook производится парсинг (выгрузка данных) данных eLibrary по статьям через API eLibrary для проекта “Паттерны коллаборации в Российском социологическом сообществе”. Осуществляется импорт необходимых библиотек, загружаются id необходимых к выгрузке статьи, производится сама выгрузка основного массива данных. Основной код выгрузки был подготовлен Львом Ципесом, в доработке кода и выгрузке данных участвовали Василиса Ващенко и Лика Капустина. Выгрузка производилась в июне-июле 2022 года. Эта версия была закончена и прокомментирована Ликой Капустиной. Представленный код можно использовать для выгрузки данных о статьях через API Elibrary. Предварительно вам нужно заключить договор с Elibrary, и тогда вы получите доступ к одному из сервисов API по IP-адресу, указанному в договоре. Для работы нужен файл формата .txt с id статей, информацию о которых нужно выгрузить через API Elibrary. Процесс выгрузки данных о статьях через API Elibrary происходит в несколько шагов:

- Шаг 1: подготовка к работе;
  - Шаг 2: выгрузка данных
  - Шаг 3: сохранение данных В результате получаются следующие файлы:
    - Промежуточные файлы с данными в формате .csv и .xlsx, которые сохраняются локально в рабочую директорию каждую 1000 итераций (с названием в формате df\_текущая дата\_число строк\_rows.формат файла);
    - final\_data.csv - pandas.DataFrame с информацией о всех статьях, находящихся в изначальном списке, в формате .csv;
    - final\_data.xlsx - pandas.DataFrame с информацией о всех статьях, находящихся в изначальном списке, в формате .xlsx;
3. Файл “2.1. Предобработка данных и получение идентифицирующих признаков.ipynb” – в этом файле проводится предобработка данных, выгруженных на этапе 1, с точки зрения работы с именами авторов и получения идентифицирующих признаков. Часть 1. Предобработка данных На данном этапе стоит задача предобработать данные, выгруженные через API Elibrary ранее (см. файл «1. Выгрузка данных по статьям»). На этом этапе стоит несколько задач:
- Предобрбатать данные по русскоязычным фамилиям и инициалам;
  - Предобрбатать данные по англоязычным фамилиям и инициалам;
  - Предобрбатать данные по русско- и англоязычным аффилиациям;
  - Провести процедуру присвоения собственных id авторов; В части 1 проводятся две первые процедуры - предобрбатка данных по русско- и англоязычным фамилиям и инициалам. Такая необходимость возникает в связи с тем, что в данных достаточно много авторов без id во внутренней системе eLibrary. Этот факт осложняет дальнейшую работу с данными и их анализ – нельзя построить сети на основании id авторов, если они не достаточно корректные. Поэтому было принято решение подготовить собственные id авторов, основываясь на ФИО и данных об аффилиациях авторов. Однако на этом этапе возникла проблема – eLibrary не контролирует единообразность записи фамилий, имен и инициалов; туда можно вписать разные знаки препинания; иногда в фамилии вписываются сразу и фамилии, и инициалы, и так далее. В этом ноутбуке демонстрируются разные проблемы и способы их решения. Получаемые в результате первой части предобработки файлы:
    - problem1.xlsx - строчки с информацией о статьях, для которых отсутствует информация даже о первом авторе (то есть, нет информации об авторах вообще). Необходимо заполнение вручную.
    - problem2.xlsx - строчки с информацией о статьях, в фамилиях и инициалах авторов которых встречаются вопросительные знаки. Необходимо заполнение вручную.
    - problem3.xlsx - строчки с информацией о статьях, среди авторов которых встречаются авторы с не-русскими фамилиями если в разделе с инициалами у них не указаны инициалы (то есть, в ячейку с фамилией отнесена и фамилия, и инициалы). Необходимо заполнение вручную.
4. Файл «2.2. Обработка аффилиаций.ipynb» – в этом файле проводится предобработка данных, выгруженных на этапе 1, с точки зрения работы с аффилиациями авторов. В этом файле проводится работа по дедупликации аффилиаций авторов путем их чистки и присуждения id. Вначале составляется общий список всех аффилиаций, затем рассматриваются варианты их

сокращения до аффилиаций и создается словарь сокращений. Работа подразумевает следующие шаги:

- Шаг 0: Предварительная очистка данных
- Шаг 1: Работа с ID аффилиаций
- Шаг 2: Работа с английским переводом аффилиаций
- Шаг 3: Создаем новые id для неидентифицированных аффилиаций
- Шаг 4: Создание новых ID для авторов
- Шаг 5: Дедупликация авторов по новым ID

5. Файл «2.3. Пост-обработка\_новых\_ID.ipynb» – в этом файле проводится пост-обработка данных – создание новых универсальных ID авторов на основе почищенных имен авторов и аффилиаций. Работа по дедупликации авторов осуществляется на основании ID авторов и аффилиаций, присужденных на предыдущих этапах 2.1 и 2.2. Для новых ID авторов были рассчитаны попарные расстояния Дамерау-Левенштайна (см. п. 5 в файле 2.2), и в этом файле новые ID анализируются на предмет пересечения важных идентифицирующих признаков для финализации дедупликации. Работа включает следующие шаги:

- Шаг 1: Обработка авторов с наличествующим идентификатором
- Шаг 2: Обработка авторов с отсутствующим идентификатором
- Шаг 3: Сбор результатов обработки воедино
- Шаг 4: Изменение рабочего файла с данными

6. Файл «“3. Анализ данных и визуализация.ipynb”» – в этом файле анализируются предварительно собранные и обработанные файлы, а также создаются графики. Работа подразумевает следующие шаги:

- Пункт 1. Распределение статей по числу авторов
- Пункт 2. Распределение цитирований
- Пункт 3. Абстракты и ключевые слова
- Пункт 4. Статистики по авторам

7. Файл «“4. Создание сетевых файлов для Pajek.ipynb”» – в этом файле создаются сетевые файлы для работы с графиками в формате Pajek. В настоящее время прописанный код подразумевает создание двумодальной сети формата «Работа – Автор», на основе которой могут строиться сети коллабораций (согласно цели проекта “Паттерны коллаборации в Российском социологическом сообществе”), однако в дальнейшем планируется написание кода для других видов сетей («Работа – Журнал», «Работа – Ключевое слово», сети соприсутствия слов в аннотациях / названиях статей и др.).

Согласно разработанной в лаборатории методологии анализа библиографических сетевых данных на русском языке дальнейший анализ сетевых данных осуществляется в программе для анализа и визуализации больших сетей Pajek.

Файлы формата Jupyter Notebook расположены в репозитории на GitHub по ссылке: <https://github.com/Daria-Maltseva/Collaboration>. Коллективом лаборатории планируется работа по продолжению развития программ ЭВМ для анализа русскоязычных данных и разработка полноценного отчуждаемого пакета для библиометрического анализа русскоязычных публикаций на языке программирования Python.

## **5 Методологические особенности предобработки данных по российским авторам в Web of Science**

Первоначальный массив данных, который лег в основу этого исследования, состоял из более чем 1.38 миллиона публикаций российских исследователей за 1990-2022 гг., индексированных в престижной международной базе данных Web of Science (WoS)[466]. Уникальность массива состоит в отсутствии любых ограничений на тип записей, количество цитирований, научную область, регион и т.д. Благодаря этому можно говорить о том, что эти данные отражают реальную картину представленности российской науки в «Web of Science» на май 2022 года, рис. 5. Исходный набор данных данного исследования включает публикации WoS, выгруженные со спецификацией поля данных «CU=(Russia)» в режиме full record (полное библиографическое описание публикаций, включающих пристатейные списки литературы). Всего исходный набор данных включал 1383996 библиографических записей о российских публикациях, проиндексированных в WoS Core Collection до мая 2022 г. Топ-20 типов документов и их количество, доля в общем объеме, цитируемость, доля в общем количестве цитирований и количество цитат на статью (CPP)

Ввиду отсутствия ограничений на этапе отбора данных и их большого размера, прежде чем приступить к анализу, мы были обязаны провести довольно крупный объем работ по предобработке данных. Сначала мы опишем общую структуру изначального массива данных, далее перейдем к процессу создания подсетов по отдельным научным категориям. Затем будут перечислены основные проблемы, которые встали перед исследовательской группой в обработке данных. После этого мы перейдем к представлению стратегии по решению конкретных проблем в именах авторов, состоящей из нескольких этапов. Наконец, будут описаны сложности, встречающиеся в записи организаций, и процесс их преодоления.

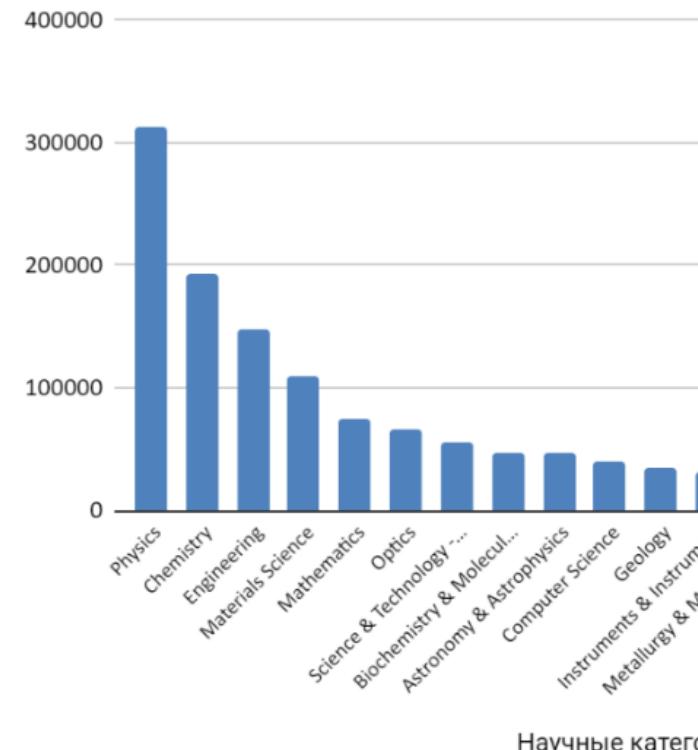
Полученные данные разделены по годам публикации, в каждой папке содержатся файлы с полными библиометрическими записями материалов в формате .txt, каждый из которых содержит максимум 500 записей. Каждая запись отделяется с помощью текстовых маркеров «PT» и «ER», как продемонстрировано на рис. 13. Основные библиометрические данные, необходимые для анализа включают в себя: название, имена авторов и их аффилиации, тип материала (статья, отчет по итогам конференции и т.д.), процитированных в работе авторов, страны, научное направление, дату публикации и уникальный идентификационный номер, присвоенный WoS.

Пример стандартной структуры полных библиометрических данных для 1 публикации

Figure 13: Пример стандартной структуры полных библиометрических данных для 1 публикации

Отбор данных по научной категории возможен с помощью параметров WC (Web of Science Categories) или SC (Research Areas): первый автоматически присваивается на основе журнала публикации, второй - определяется самими авторами, а потому обладает более точным определением научных направлений. Всего по SC в массиве определяется около 156 научных направлений, из них на первые 20

## ТОП-20 научный категорий по количеству публикаций



категорий приходится 68% всех публикаций (рис. ??).

Подмножество по социологии было выделено на основе категории SC как категории исследования WoS (выполнение условия « $SC = Sociology$ ») и состоит из 7915 публикаций (менее 0,01% от всего массива публикаций). Для большинства публикаций в Web of Science характерна принадлежность нескольким категориям исследования. В исходном наборе данных у некоторых публикаций насчитывалось до 9 таких категорий, в среднем для публикации характерно 3-4 категории исследования. В нашем случае в качестве материалов исследования были отобраны все публикации, у которых как минимум одна категория исследования была указана как Sociology. Подмножество по социологии включает в себя публикации всех представленных типов – статьи, главы в монографиях, конференционные материалы и т.д. – за период с 1992 до мая 2022 года.

Предобработка данных была реализована в Python с дополнительной ручной проверкой и корректировкой. Особое внимание уделялось именно авторам и организациям для приведения к единому виду и объединению разных вариантов написания фамилий и имен авторов, а также разных вариантов применения названий организаций. В целом на данном этапе было выявлено достаточно большое количество сложностей; некоторые из них приведены ниже как категории особенностей предобработки данных:

1. транслитерация ФИО авторов с кириллицы на латиницу (французский и английский варианты написания, сложные звуки (шипящие)). Пример: loukov – lukov, toshchenko – toshenko, alikhadzhieva – alikhadjieva и т.д.;
2. использование у российских авторов отчества, которое в библиографическом описании публикации может как присутствовать, так и отсутствовать. Отчество может стоять на первом месте вместо фамилии (valerevich, radaev vadim) или отсутствовать в принципе (radaev, vadim), что мультилинирует количество авторов. В этом случае усложняется обработка данных, так как для

- одного канонического написания ФИО автора (radaev, v. v.) необходимо выявить все разнообразные варианты имени этого автора в наборе данных. Например, для канонического ФИО автора как toshchenko, i. z. было выявлено 13 разных вариантов написания;
3. разные написания названий организаций – название организации с советского периода (Томский государственный университет им. В.В.Куйбышева – в настоящее время Национальный исследовательский Томский государственный университет), множественные варианты аффилиаций (Высшая школы экономики, НИУ Высшая школы экономики, Национальный исследовательский университет и т.д.);
  4. сложная организационная структура (несколько кампусов, множество институтов и т.д.), когда указанное подразделение кодируется WoS как самостоятельная организация;
  5. в библиографических описаниях публикаций в WoS категория независимого исследователя (independent researcher) никаким образом не учитывается, рассматривается как пропущенное значение. Также по непонятным причинам, в WoS не были указаны ряд аффилиаций коммерческих организаций, названия которых в публикациях были написаны по всем правилам (Yandex, GetBrand, Издательский дом «Коммерсант» и др.);
  6. технические ошибки, опечатки и пропущенные данные (особенно у старых публикаций).

Предобработка данных об авторах включала два основных процесса: (1) предобработка данные по определенным правилам; (2) поиск схожих авторов, чье ФИО отличается на несколько символов с использованием технологии fuzzy matching. Главным приоритетом в работе, помимо максимального снижения числа «универсальных» ФИО авторов (в данных о российских социологах встречается до 8 вариаций ФИО одних и тех же авторов) являлось также минимальное число некорректных случаев соответствия ФИО авторов. Здесь и далее “ФИО авторов” и “имена авторов” являются синонимичными понятиями.

Одной из главных проблем, решенных на первом этапе работы с авторами стала проблема не унифицированности записей об авторах: хотя большинство записей имело вид “, <отчество или первая буква отчества, если указано>”, но встречались и другие формы записи, когда на месте фамилии находилось отчество или имя. Нам был реализован алгоритм поиска имен и отчеств на некорректном месте и изменения порядка записей в сторону унифицированных (*andrey, kinyakin -> kinyakin, andrey; sergey, stepanov -> stepanov, sergey*). Помимо этого, мы заменили мало популярные формы имен на популярные (“*nadejda*”-> “*nadezhda*”), убрали лишние символы из имен (например, символ штриха), чтобы облегчить поиск сильно совпадающих имен авторов на следующем этапе.

После первого этапа предобработки мы имели предварительно обработанные имена авторов, приведенные к единому формату записи: “, <отчество или первая буква отчества, если указано>”. После этого мы решили реализовать процесс мэтчинга авторов на основании метрики расстояния Левенштейна: эта метрика позволяет получить число символов, на которые отличаются определенные строки. При простом поиске совпадающих строк с именами авторов существует риск случайно слить в один разных в реальности авторов; например, хотя для пары “*barsukova, s y*” и “*barsyukova, s y*” расстояние Левенштейна будет равно единице и этот мэтч может быть обработан далее корректно (это действительно один и тот же автор), то при отсутствии дополнительных правил мы могли бы привести к единой форме в реальности разных авторов, например, расстояние Левенштейна для строк “*ivanova, a a*”

и “ivanov, a a” также равно единице. Поэтому мы реализовали поиск совпадающих авторов при наличии дополнительных правил: совпадение последнего звука фамилии (чтобы исключить мэтчинг мужчин и женщин-авторок), совпадение первого звука фамилии, совпадение всех или хотя бы части инициалов в том случае, если отчество автора не было указано. Далее производилось деление случаев на категории по типам (гласные/согласные звуки) для первых отличающихся звуков в фамилиях и совпадении этих типов:

Тип	Уверенность в корректности дальнейшего мэтчинга фамилий   Пример
:—:	:—:   :—:
Фамилии отличаются на гласные звуки	Высокая   alekseeva, t. a. – alekseyeva, t. a.
Фамилии отличаются на согласные звуки	Средняя   sebentsov, a. b. – sebentzov, a. b.
Фамилии, отличающиеся лишь из-за одного дублирующегося звука	Высокая   isaev, l. m. – issaev, l. m.
Фамилии отличаются на гласный и согласный звук	Низкая   lapin, v. s. –apkin, v.  **

Далее, во время формулирования разных категорий потенциальных мэтчей, возможно также посмотреть содержание всех списков и далее вручную удалить некорректные пары, однако, можно использовать и целый объект, полученный на этом шаге. Таким образом, на данных по российским социологам после проведения процедур первого типа удалось сократить число уникальных авторов на 11,2%, после проведения мэтчинга на основе поиска совпадающих фамилий еще на 4,8% от изначального числа уникальных авторов, а итоговый результат составил 16%-ное снижение уникальных авторов. По ощущениям, появившимся при первичном просмотре файла по социологам, примерно 20-25% всех записей являлись дублями и могли бы быть приведены к другой, более популярной форме имени автора. Полученная нами цифра, с одной стороны, не очень маленькая – что демонстрирует, что данный, даже очень аккуратный подход к мэтчингу авторов, способен снижать число уникальных авторов; с другой стороны, это относительно невысокий показатель, демонстрирующий аккуратность подхода авторов к поиску совпадающих авторов. Описанный процесс выше необходим для дальнейшего построения сетевых моделей на основе данных об авторстве и соавторстве, так как при наличии большого количества дублирующихся записей об авторах выводы любого исследования будут некорректны. Дальнейшие планы по развитию проекта связаны с намерением повысить точность мэтчей и составить новые процедуры поиска совпадающих имен авторов.

Деятельность по предобработке данных об организациях, с которыми аффилированы авторы исследуемых публикаций, включала в себя два ключевых составных блока: итеративный fuzzy matching и следующий за ним keyword matching.

На предварительном этапе обработки данных организаций были определены проблемные аффилиации, для которых вместо названия указан адрес, и исключены из последующего анализа. К таковым были отнесены строки, содержащие цифры или маркеры адреса, такие как слова «lane» (переулок), «str» (ул.) и др. Эти аффилиации были размечены вручную при помощи обращения к публикациям, к которым они относятся. Прочие аффилиации были очищены от специальных символов и отдельно стоящих букв, а затем разделены на слова-токены.

После предобработки, мы приступили к итеративному мэтчингу токенов на основании метрики близость Дамерау-Левенштайна, реализованной в библиотеке jellyfish для языка программирования Python. Сначала мэтчинг применялся к отдельным токенам: для всех пар токенов длиннее 3 букв

рассчитывалось расстояние Дамерау-Левенштайна и, после ручной проверки, пары с расстоянием менее 3 (токены отличаются друг от друга менее, чем 3 символами) были объединены. Подобная операция была произведена три раза последовательно с сокращением порога объединения до 1 отличающегося символа. Это позволило объединить слова с разным написанием при транслитерации на английский (e.g. ‘altay’ и ‘altai’), альтернативные сокращения (e.g. ‘federal’ и ‘federat’), опечатки (e.g. ‘novasibirsk’ и ‘novosibirsk’), имена (e.g. ‘peter’ и ‘petr’), а также слова, написание которых варьируется между языками написания (e.g. ‘milan’ и ‘milano’ или ‘labor’ и ‘labour’). Затем дедуплицированные токены были снова объединены в полные названия. К измененным строкам названий также был применен мэтчинг: объединялись строки, отличающиеся не более чем на 2 символа.

Дедупликация токенов и мэтчинг строк позволяют избавиться лишь от части вариативности в написании названий организаций ввиду того, что наименования могут включать в себя слова в разной последовательности, неоднородный перевод, сокращения, разную степень детализации аффилиации (например, до уровня факультета). Дальнейшее удаление дубликатов производилось при помощи подхода, основанного на выделении ключевых слов для идентификации ряда крупных организаций и присвоении стандартизованных названий всем наблюдениям, содержащим указанные ключевые слова. Полный список использованных ключевых слов представлен в таблице ниже. При выделении ключевых слов мы ориентировались на задачу обнаружения последовательности минимальной длины, которая позволяет обнаружить как можно большее количество строк, относящихся к искомой организации.

Организация	Ключевые слова
НИУ ВШЭ	hse, higher_sch, higher_econ
МГУ им. Ломоносова	lomonosov, msu
МГТУ им. Баумана	bauman
Российская Академия Наук	russian_acad_sci, ras
РУДН	friend, rudn
РАНХиГС	ranepa, russian_acad_natl_econ_publ
РГГУ им. Плеханова	plek
МГИМО	mgimo, inst_int_relat

На этом этапе достигается наибольшее падение в количестве уникальных аффилиаций в базе данных. Финальным штрихом в автоматизированной обработке организаций стало приведение к однородному написанию всех государственных организаций и, в частности, министерств. После обработки все аффилиации с министерствами записываются в однородном формате: строки начинаются с ‘russian\_minist’. Это также позволило идентифицировать и устраниить ряд дубликатов.

Дальнейшая работа с аффилиациями требовала экспертного вмешательства. Так, были идентифицированы «подозрительные» аффилиации, которые были затем проанализированы вручную. К «подозрительным» были отнесены аффилиации, длина которых не превышает 5 символов, а также содержащие слова «faculty», «fac», «dept», «school», «inst» с целью обнаружения случаев, в которых в аффилиации сохранилось только подразделение, а не основная организация. Аналогично, ручной обработки требовало сопоставление не англоязычных аффилиаций с англоязычными: в нашей базе данных присутствуют названия организаций не только на английском, но и на испанском, итальянском

и немецком.

По итогам обработки удалось сократить количество уникальных аффилиаций в выборке с 1644 до 1309 (на 21%).

## 6 Сравнительный анализ возможностей баз данных Web of Science и eLibrary для анализа библиографических сетей

### 6.1 Введение

В статье проводится сравнительный анализ баз данных научных публикаций Web of Science Core Collection и eLibrary с целью выделения их особенностей и возможностей для анализа библиографических сетей российских авторов. Актуальность исследования определяется необходимостью адаптации и разработки подходов и инструментов для сбора, предобработки и анализа библиографических данных на русском языке.

Анализ библиографических сетей – частный случай применения методологии анализа социальных сетей. Он основан на построении и анализе сетей соавторства и коллаборации, цитирования и социтирования, библиографического сочленения, соприсутствия библиометрических единиц анализа. Направление способно показать закономерности развития взаимодействия в научном сообществе, определить его структуру, динамику, направления исследований [bar-ilan2008b?, mingers2015b?, rousseau2018b?]. Основные этапы исследования с применением анализа библиографических сетей подразумевают использование технологических решений для 1) формирования базы библиографических данных, 2) ее предобработки и построения различных видов библиографических сетей и 3) последующего изучения с применением методов сетевого анализа (social network analysis). Как и в любом исследовании, выбор источника информации является определяющим для качества анализа – по принципу GIGO (“garbage in, garbage out”), получение достоверных результатов напрямую зависят от стратегии поиска и полноты используемой базы данных. Выбор баз данных для исследователя является достаточно широким – помимо часто используемых для учета эффективности работы ученых баз научного цитирования Web of Science (WoS) и Scopus, большую популярность приобрели бесплатные базы данных, агрегирующие библиографическую информацию, такие как “универсальные” Google Scholar и OpenAlex, включающие информацию о патентах Digital Science Dimensions и Lens, базы медицинских исследований PubMed и Cochrane, научные социальные медиа SciFinder, Mendeley, и др. Эти базы часто выступают источниками данных в библиометрических исследованиях, где используются специально разработанные алгоритмы для сбора данных с этих площадок и их анализа.

Так, пакеты Bibliometrix для R и Python и их веб-приложение Biblioshiny [aria2017a?] позволяют работать с базами данных Scopus, WoS, PubMed, Digital Science Dimensions, Cochrane, Lens и OpenAlex для анализа со-цитирования, библиографического сочленения библиографических единиц анализа, соавторства и со-присутствия ключевых слов. Программа VOSviewer, помимо перечисленных, позволяет работать с такими базами как Crossref, Europe PMC, Semantic Scholar, OpenCitations, и WikiData через их API-сервисы, запрашиваемые в интерактивном режиме в самой программе. Программа CitNetExplorer, предназначенная для анализа цитирований научной литературы, импортирует данные из WoS. На использование данных WoS ориентирован и методологический подход, разработанный В.Батагелем, А.

Ферлигой и П. Дореаном, применявшимся для анализа некоторых зарубежных научных дисциплин и описанный в отечественной литературе [9], который использует специально разработанную программу WoS2Pajek для создания из данных WoS сетевых файлов, пригодных для работы в программе для анализа и визуализации больших сетей Pajek. Значительное количество библиотек научных публикаций привело к появлению исследований, посвященных сравнительному анализу различных площадок, где они сравниваются по различным характеристикам – покрытию по времени, дисциплинам, странам, журналам и типам публикаций, их пересечению, формату и глубине метаданных библиографических описаний, удобству использования и выгрузки информации [bar-ilan2008b?, mingers2015b?, rousseau2018b?].

При изучении данных русскоязычных авторов исследователь сталкивается с проблемами покрытия данных и методологии их анализа. С точки зрения покрытия, представленность работ на русском языке в международных базах научных публикаций зависит от включенности отечественных журналов в списки индексируемых источников и может быть характеризована как частичная. Учет научной продуктивности только на основе международных баз данных существенно занижает объем производимой в России научной литературы. Однако попытки учета отечественных публикаций из российских источников сталкиваются с другой проблемой – богатый инструментарий, разработанный в области библиометрического анализа, ориентирован на публикации, написанные на английском языке. Заложенные в существующих инструментах алгоритмы предобработки данных (дизамбигуации имен, лемматизации, токенизации слов и т. д.) ориентированы на англоязычные коллекции словарей и не могут быть напрямую использованы для нормализации данных на русском языке. Следовательно, эти инструменты могут быть использованы для публикаций российских авторов, написанных на английском языке, однако при анализе данных на русском языке возникает задача по адаптации существующих или разработке новых подходов и инструментов библиометрического анализа.

В поле российской науки имеется несколько баз данных, аккумулирующих информацию о научных публикациях. Крупнейшей в России базой научных публикаций является научная электронная библиотека eLibrary, которая интегрирована с Российским индексом научного цитирования (РИНЦ) – созданным по заказу Минобрнауки РФ общедоступным инструментом измерения публикационной активности ученых и организаций. В качестве альтернативы можно отметить научную библиотеку КиберЛенинка, предоставляющую доступ к публикациям на основе принципов открытой науки. Преимущество eLibrary заключается в том, что она предлагает встроенные алгоритмы для анализа публикационной активности авторов и организаций; однако никаких инструментов для сетевого библиометрического анализа в базе нет. Полноценной методологии по сбору, предобработке и анализу библиографических данных на русском языке до настоящего времени не представлено. В связи с этим методика библиометрического сетевого анализа практически не используется при изучении российской науки, за редким исключением [452, 477, 479].

Данная статья носит методологический характер и нацелена на сравнительный анализ возможностей баз данных WoS и eLibrary для анализа библиографических сетей российских авторов. В качестве примера взяты все публикации в области социологии за 2010-2021 гг. на обеих площадках – 3,995 публикаций в WoS и 75,232 публикаций в eLibrary. В результате литературного обзора выделяются основания для сравнения двух баз данных, содержащие различные параметры, которые затем анализируются посредством описательного, статистического и сетевого анализа для поиска пересечений между массивами данных и содержательными результатами анализа. Анализ составляющих

их параметров позволяет найти пересечения между массивами данных и содержательными результатами анализа. Делаются выводы о соотношении двух баз, их возможностях и ограничениях по использованию в качестве основного (единственного) источника информации, даются рекомендации об их использовании для изучения отечественной науки.

## 6.2 Обзор литературы

Платформа WoS компании Clarivate Analytics является первой базой научного цитирования, построенной на основе Индекса цитирования научных статей (Science Citation Index), разработанного в 1960-е гг. одним из основателей наукометрии Ю.Гарфилдом. На основе анализа цитирований статей Гарфилд разработал подход к рейтингованию научных журналов, составляющих “ядро” научных дисциплин. Позже в “ядре” (Core Collection - CC) WoS появились также индексы цитирования социальных наук (Social Sciences Citation Index - SSCI), искусств и гуманитарных наук (Arts and Humanities Citation Index - AHCI) и новых источников (Emerging Sources Citation Index - ESCI). Дополнительно в WoS входят индексы цитирований конференционных публикаций и книг, а также национальные индексы цитирований, такие как Russian Science Citation Index (RSCI). Долгое время WoS имела монополию на предоставление информации о научной литературе, однако с появлением в 2004 г. платформ Scopus и Google Scholar ситуация изменилась. Scopus, как и WoS, стал предоставлять информацию о цитировании, получаемую в виде метаданных от производителей научной литературы, однако существенно расширил покрытие научных журналов. В то же время Google Scholar расширил диапазон источников до материалов конференций, книг, диссертаций, отчетов и других типов публикаций с сайтов издателей и конференций, используя автоматические методы извлечения информации из электронных файлов научных публикаций,

В реview о развитии наукометрических исследований [bar-ilan2008b?, mingers2015b?] приводится масса ссылок на исследования, сравнивающие базы данных WoS, Scopus и Google Scholar друг с другом, а также с другими базами по научным дисциплинам (MEDLINE, CiteSeer, ResearchIndex и др.). С точки зрения дизайна исследований, авторы делают выгрузки по определенной теме или журналам из разных баз и сравнивают их по покрытию (количеству найденных публикаций) и распределению публикаций по дисциплинам и областям наук, типам документов, странам, языкам написания, а также количеству полученных внутри базы цитирований. Сравнения показали наличие существенных различий между базами по охвату научных дисциплин: тогда как Google Scholar обеспечивает широкий охват большинства предметных областей, WoS и Scopus имеют меньшие публикаций в целом, хорошо охватывают естественные науки, умеренно представлены в области социальных наук и слабо - в области искусства и гуманитарных наук [harzing2016b?]. В связи с особенностями формирования баз данных количество работ по всем дисциплинам в Scopus выше, чем в WoS (большее количество журналов); однако качество метаданных в обеих базах существенно выше, чем в Google Scholar (автоматическая экстракция данных). Важной особенностью WoS является охват библиографических публикаций вплоть до 1900 года, включая списки цитируемой литературы (важная информация для анализа цитирований). В Scopus информация о цитировании включена в описания публикаций только с 1970 г., в Google Scholar она отсутствует.

С 2004 г. появилось много других агрегаторов научной информации, такие как OpenAlex (универсальная база), Digital Science Dimensions и Lens (начинались как патентные базы), PubMed и Cochrane (медицинские исследования), SciFinder, Mendeley, ResearchGate (научные социальные медиа)

и др., каждый из которых имеет свои возможности и ограничения. Недавние работы, сравнивающие “традиционные” базы с новыми – Microsoft Academic (2016), CrossRef (2017), Dimensions (2018), COCI – OpenCitations Index of Crossref (2019) (например [157, 355, martin-martin2021a?], см. также рассматриваемые ими публикации) – показывают, что Google Scholar по-прежнему остается наиболее полным источником, а Microsoft Academic и Dimensions являются хорошей альтернативой Scopus и WoS с точки зрения охвата. Однако есть основания утверждать, что WoS и Scopus все равно остаются самыми популярными источниками информации для научометрических исследований [441].

Крупнейшей базой научных публикаций в России, а также научной периодики на русском языке в мире, является научная электронная библиотека eLibrary. eLibrary интегрирована с Российским индексом научного цитирования (РИНЦ), и через информационно-аналитическую систему SCIENCE INDEX позволяет измерять публикационную активность ученых и организаций. С 2016 г. публикации из лучших российских журналов по всем научным направлениям из РИНЦ (порядка 600-700 журналов, все выпуски за последние 10 лет) индексируются на платформе Web of Science в виде отдельной базы данных Russian Science Citation Index (RSCI). Эта база RSCI, вместе с публикациями российских авторов, индексируемыми в Scopus и Web of Science, входит в т.н. “ядро РИНЦ”. Три базы имеют частично пересекающиеся коллекции журналов. В исследовании [146] приводятся данные о том, что в 2019 г. в Scopus входило 488 российских журналов, в Web of Science - 353, а в RSCI - 650. С 2016 г. количество журналов, общих для RSCI и Scopus, увеличилось с 89 до 205, общих для RSCI и WoS - с 8 до 105; а общих для трех баз - с 5 до 74. Формирование базы RSCI сопряжено с критическими оценками [kassian2019b?], которые относятся к процессу отбора и качеству включенных в базу журналов.

В исследовании, нацеленном на сравнение коллекций публикаций российских авторов в базах WoS и Scopus (без базы RSCI), был показан экспоненциальный рост числа публикаций с 2006 по 2016 гг. [270]. Исследователи показали, что динамика по количеству публикаций зависит от используемой базы данных и изменений в ее охвате (например, использование русского как публикационного языка привело к росту доли публикаций с 4.8% до 14.8% в 2006 - 2016 гг. в Scopus). Сравнение двух баз с RSCI<sup>1</sup>, проведенное аналитиками eLibrary [458], позволяет сделать некоторые выводы о пересечении между разными коллекциями публикаций российских авторов. В 2018 г. количество публикаций в Scopus составило 94.3 тыс. работ, из них 43.7 тыс. работ не пересекались с квартилями Q1-Q4 в WoS. Количество публикаций российских авторов в WoS CC составило 71.8 тыс. работ, включая 49.4 тыс. работ в журналах, входящих в квартили Q1-Q4, и 20.9 тыс. работ, индексируемых в ESCI. Число работ в базе RSCI в 2018 г. составило 72.5 тыс. публикаций, в т.ч. 69.8 тыс. работ в журналах, не входящих в квартили Q1-Q4 в WOS CC (т.е. входили в Q1-Q4 только 2.7 тыс. публикаций). Таким образом, база RSCI имеет достаточно большое количество уникальных публикаций и ее вклад в ядро РИНЦ является значительным. Анализика [458] показывает, что средние цитирования для статей российских авторов, опубликованных в WoS Q1-Q4, составляют соответственно 51.07, 25.55, 16.12 и 7.93, а в Scopus Q1-Q4 - соответственно 37.51, 10.60, 4.68 и 2.93. Средняя цитируемость российских статей в RSCI (2.51) сравнима с Q-4 Scopus и заметно выше, чем в ESCI (1.81).

Исследователи [mingers2015b?] объясняют ситуацию с более низким покрытием социальных и гуманитарных наук в базах научного цитирования тем, что во многих дисциплинах журналы являются

<sup>1</sup>Работ, нацеленных на сравнение баз РИНЦ и RSCI с другими зарубежными базами научного цитирования, в рамках проведенного поиска обнаружено не было.

не единственными престижными источниками обмена научными знаниями; важными могут быть также монографии, материалы конференций или отчеты, ориентированные на широкую общественность. Помимо научных публикаций, важными видами научного вклада могут быть также программное обеспечение, патенты и т. д. Для различных наук характерна своя публикационная культура: исследователи отмечают, что в социальных науках часто встречаются «ученые-одиночки», тогда как работа в естественных и экспериментальных науках чаще осуществляется командами исследователей, которым легче стать “видимыми” за счет большего количества публикаций и цитирований. Как результат, в ряде работ обсуждаются проблемы покрытия и необходимости использования нескольких баз данных для формирования наиболее полного массива исследования. Обзорные работы также показывают, что исследования, нацеленные на изучение развития наук в конкретных странах, сравнивают массивы публикаций из международных и национальных баз данных научного цитирования, а также поднимают специфические вопросы репрезентации данных, например, на других языках. Так, исследователи выявили, что около 50% имен на испанском языке имеют несколько вариаций написания даже внутри одной и той же международной базы [336]. В научной литературе также рассматриваются техники и проблемы, связанные с выгрузкой данных из различных библиографических агрегаторов.

Говоря о российских публикациях, исследователи [270] пришли к выводу, что база WoS, и особенно Scopus, должны с осторожностью применяться в качестве единственных инструментов измерения результативности российских исследователей. Систематический поиск научной литературы показывает, что в настоящее время РИНЦ и RSCI<sup>2</sup> выступают источниками данных в библиометрических исследованиях, ориентированных на изучение российского научного поля. Однако методология сбора таких данных для баз РИНЦ и RSCI при этом не представлена.

### 6.2.1 Методология и данные исследования

**6.2.1.1 Методология** На основе рассмотренных работ, сопоставляющих базы данных научных публикаций, а также предварительного анализа рассматриваемых массивов, были сформулированы основания и параметры для сравнения баз WoS и eLibrary (Табл. ??). Тогда как сравнение процесса работы с данными осуществлялось в рамках описательного анализа доступной информации, дизайн исследования для более детального анализа подразумевал сравнение аналогичных массивов данных, выгруженных из каждой базы (согласно дизайном исследований, рассмотренных в обзоре). Два массива рассматривались по определенным параметрам с помощью статистического анализа; затем на основе данных из массива было построено несколько базовых библиометрических сетей, которые также рассматривались в соотношении друг с другом по ряду рассчитанных параметров.

Анализ указанных параметров позволяет найти пересечения между массивами с точки зрения присутствующих в них единиц анализа – публикаций, журналов, авторов, ключевых слов – и оценить размер множеств, находящихся на пересечении и при объединении массивов. Помимо этого, сравнение результатов анализа с точки зрения содержания также позволяет сделать выводы о том, насколько похожие результаты дает использование двух баз данных.

---

<sup>2</sup>Не всегда можно понять, какая база использовалась, т.к. на английский язык “Российский индекс научного цитирования” переводят не только как “Russian Index of Science Citation”, но и как “Russian Science Citation Index” (повторяет название в базе WoS). Однако, как видно из обзора, РИНЦ и RSCI – сходные, но не одинаковые базы.

Table 5: Основания, параметры и способы анализа для сравнения баз данных

Основания	Параметры	Способы анализа
Процесс работы с данными	- особенности сбора данных, - формат и структура получаемых массивов, - количество используемых в библиографических описаниях метаданных, - возможности предобработки данных существующим программным обеспечением, - возможности построения файлов для сетевого библиометрического анализа	Описательный анализ
Массивы данных	- размер массивов (число публикаций), - динамика числа публикаций во времени, - объем пропущенных значений по метаданным в библиографических описаниях публикаций, - количество уникальных библиометрических единиц (публикаций, авторов, журналов, ключевых слов)	Статистический анализ
Производные сети	- распределение уникальных авторов и ключевых слов по частоте встречаемости в публикациях в массиве, - распределение соавторов по авторам в массиве, - топ наиболее частотных журналов, ключевых слов, - топ авторов по числу работ и числу соавторов	Статистический и сетевой анализ двумодальных сетей работ и авторов WA, работ и ключевых слов WK, одномодальной сети коллабораций Co.

С точки зрения методологии анализа данных, проводимого для количественной оценки по выделенным основаниям, использовались инструменты статистического и сетевого анализа. Существуют разные возможности работы с библиометрическими данными из базы WoS, которые имплементированы на специализированных площадках (например, VOSviewer, InCite), а также реализованы в специальных пакетах программ для статистического анализа данных (R, Python). В исследовании подсчет общей статистики и сетевых характеристик по набору данных WoS проводился с помощью пакета Bibliometrix в

R и его приложения Biblioshiny, автоматическичитывающих загружаемый набор данных, и программы для сетевого анализа Pajek, требующей предварительной подготовки файлов для анализа через программу WoS2Pajek<sup>3</sup>. Статистический анализ также проводился с помощью отдельных библиотек в Python и R и программы Microsoft Excel. Данные eLibrary представлены в другом формате и на русском языке, поэтому указанные программы не могли быть напрямую к ним применены. Использовался разработанный авторами статьи методологический подход, включающий сбор, предобработку данных и построение сетевых файлов, который подразумевает использование программ Python и Pajek. В рамках представления результатов анализа описываются особенности работы в данных программах и сравниваются их аналитические возможности.

**6.2.1.2 *Данные*** В статье сравниваются публикации российских социологов на площадках WoS и eLibrary. Единицами анализа являются научные статьи в научных журналах в области социологии. Оба массива данных включают публикации за 2010–2021 гг. Ниже представлена информация о размере массивов, деталях сбора и формата итоговых данных.

**Размеры массивов данных.** В массив данных eLibrary входит 75 232 научные статьи из научных журналов, представленных на сайте eLibrary (имеющих заключенный договор). Анализируемый массив данных собран в рамках исследовательского проекта, выполняемого в рамках гранта РНФ<sup>4</sup>. В массив данных WoS входит 3,559 научные публикации типа «article» в научных журналах, индексируемых в базе WoS CC<sup>5</sup>. Анализируемый массив данных является подмножеством из набора данных, собранных в рамках проекта по изучению российской науки на основе публикаций, представленных в WoS CC (1,383,996 библиографических записей о российских публикациях за период с 1992 до мая 2022 г.).

**Стратегия сбора данных.** При сборе данных из eLibrary работа проводилась совместно с сотрудниками ООО “Научная электронная библиотека”, осуществляющими поддержку этой базы. На первом этапе из всех работ в журналах, представленных в eLibrary, относящихся по ГРНТИ к области «Социология», были отфильтрованы статьи, где по крайней мере одним из авторов является российский ученый (в поле страны указана Россия). По данному запросу был составлен список из 75 232 уникальных идентификаторов публикаций, по которым затем была собрана полная информация.

Стратегия сбора данных из WoS подразумевала использование базы Core Collection. Были отобраны и выгружены все статьи российских авторов (в поле страны указана Россия, «CU=Russia»). Подмножество по социологии было выделено на основе категории (“research area”), к которой относится публикация (поле «SC=Sociology»). Первоначально массив состоял из 7,915 публикаций (35 из которых были дублирующими), но для целей настоящего анализа он был ограничен по типу публикаций (поле «DT = Article») и временному периоду (2010-2021).

**Сбор данных.** Платформа eLibrary и база РИНЦ не подразумевают функционала выгрузки данных в каком-либо формате библиографических описаний (в отличие от других баз данных). Поэтому сбор библиографических описаний из eLibrary осуществлялся через API-сервис, доступ к которому был предоставлен в рамках договора с ООО “НЭБ”. Используя полученный список из идентификаторов

<sup>3</sup>Особенности работы с разными программами при анализе одного набора данных подробно описываются в статье этого номера – Павлова И.А. и др. Международные научные коллaborации российского социологического сообщества.

<sup>4</sup>Проект «Паттерны коллаборации в российском социологическом сообществе: структура научных школ и возможные точки роста» выполняется в рамках гранта Российского научного фонда в 2021-2023 г. под руководством Д.В. Мальцевой.

<sup>5</sup>Проект осуществляется совместно Д. Фиалой, отвечающим за сбор и исследовательский анализ данных, и коллективом МЛ ПСА под руководством Д.В. Мальцевой, отвечающим за сетевой анализ массива.

публикаций, через соответствующий сервис API делались запросы на информацию по каждой публикации. Выдача данных представляет собой XML-страницу структурированного вида с идентифицированными полями (см. пример на Рис.1). Выгрузка данных осуществлялась из нужных полей, а затем записывалась в единый файл с помощью языка программирования Python<sup>6</sup>. Данные собраны в октябре 2022 г.

Платформа WoS дает возможность выгрузки библиографических описаний отобранных работ в различных форматах (RIS, Excel, обычный текстовый файл - plain text) для всех зарегистрированных пользователей (чей доступ осуществляется через организационную подписку). Используемый формат представляет собой текстовый файл структурированного вида с идентифицированными полями (Рис. 6.2.1.2). В зависимости от того собирается или нет информация о цитируемой литературе (поле “CR”), за одну итерацию можно загрузить до 500 или 1000 библиографических описаний в едином файле. Для сбора данных был написан программный код на Python, который позволял итеративно обращаться к базе и последовательно собирать файлы с описаниями. Данные собраны в мае 2022 г.

Пример библиографических описаний публикаций: eLibrary (XML-страница, сверху) и WoS (текстовый файл, снизу) Пример библиографических описаний публикаций: eLibrary (XML-страница, сверху) и WoS (текстовый файл, снизу)

**Формат и структура данных.** Массив данных eLibrary представлен в виде таблицы в формате .csv, в которой приведена информация по всем полям, доступным к выгрузке, для всех 75,232 публикаций. Данные WoS CC представлены в виде единого текстового файла в формате .txt с полным библиографическим описанием 3,559 публикаций, включая пристатейные списки литературы. В обоих массивах содержится информация по таким библиографическим единицам как публикация, автор(ы) и журнал. Набор метаданных, которыми описываются библиографические единицы в каждом массиве, приведен в Табл. ??.

В целом можно видеть, что базы похожи по используемым ими метаданным. Очевидное отличие базы WoS CC заключается в том, что в ней не представлена информация на русском языке (название, аннотация, ключевые слова, имя автора). В этой базе также отсутствует информация об ID авторов и их организаций, ID журналов и издательства – хотя она может дать важные идентифицирующие признаки при решении проблемы дизамбигуации единиц анализа (хотя нужно отметить, что для журналов WoS помимо полного названия приводят и два вида его стандартизированной аббревиатуры, что может быть использовано для обозначенной цели). В этом смысле наличие ID в описаниях eLibrary выгодно отличает эту базу и предоставляет исследователям дополнительные аналитические возможности. В данных eLibrary также указываются другие базы, в которые входит публикация (РИНЦ, RSCI, WoS, Scopus и BAK); в WoS предоставляется информация только о базе внутри ядра CC, к которой принадлежит публикация.

Главным выгодным отличием базы WoS CC является наличие информации о цитируемой литературе (поле “CR”), что позволяет проводить определенные виды анализа - изучать сети цитирований, со-цитирований, библиографического сочленения между различными библиографическими единицами (авторами, публикациями, журналами, и т.д.). В обеих базах подсчитывается также число цитирований, полученных внутри данной базы и других баз. Помимо цитирования, в WoS есть также показатель по использованию публикации другими авторами за определенные периоды времени (доступу к тексту и загрузке), что также показывает внешний интерес к научной работе.

---

<sup>6</sup>Код для выгрузки статей, а также их предобработки и построения сетевых файлов для последующего анализа в настоящее время регистрируется в качестве результата интеллектуальной деятельности и доступен в репозитории на GitHub.

Table 6: Метаданные библиографических описаний в WoS и РИНЦ

№	Параметр	WoS	РИНЦ
<b>Публикация</b>			
1	ID публикации	+	+
2	Название на русском языке	-	+
3	Название на английском языке	+	+
4	DOI публикации	+	+
5	Дата публикации (год)	+	+
6	Предметная область	+	+
7	Язык публикации	+	-
8	Тип публикации	+	-
9	Количество страниц (начальная и конечная страницы)	+	+
10	Число цитирований	+	+
11	Число использований (доступ, скачивание)	+	-
12	Аннотация на русском языке	-	+
13	Аннотация на английском языке	+	+
14	Ключевые слова на русском языке	-	+
15	Ключевые слова на английском языке	+	+
16	Информация о финансировании	-	+
17	Гиперссылка на статью в базе	-	+
18	Библиографическое описание	-	+
19	Список цитируемой литературы	+	-
20	Количество процитированной литературы	+	-
<b>Автор(ы)</b>			
1	Фамилия и инициалы на русском языке	-	+
2	Фамилия и инициалы на английском языке	+	+
3	ID автора	-	+
4	Название аффилиации автора	+	+
5	ID аффилиации автора	-	+
6	Местоположение (страна, город)	+	-
<b>Журнал</b>			
1	Название журнала	+	+
2	ID журнала	-	+
3	ISSN / e-ISSN	+	+
4	Импакт-фактор	-	+
5	Включенность в другие базы данных	+-	+
6	Выпуск	+	+
7	Номер	+	+
8	Название издательства	+	+

№	Параметр	WoS	РИНЦ
9	ID издательства	-	+
10	Адрес издательства (город и почтовый адрес)	+	-

**Предобработка данных.** Поскольку логика предобработки данных массива eLibrary отчасти следует логике работы с данными WoS, вначале рассмотрим этот набор данных. Предварительный исследовательский анализ данных массива WoS в программе Biblioshiny (см. Табл. ?? ниже) показал, что значительная часть данных отсутствует в полях аннотаций и ключевых слов (поле “DE”, Keywords) – около 21% – и полях с дополнительными ключевыми словами (поле “ID”, Keywords Plus) и DOI – 50% и 79% соответственно. Предобработка данных для массива WoS проводилась с помощью программы WoS2Pajek (URL: <http://vladowiki.fmf.uni-lj.si/doku.php?id=paiek:wos2paiek>), используемую для построения сетевых файлов. Рабочее окно программы показано на Рис. 2. Функционал программы подразумевает чистку исходного файла с библиографическими описаниями - автоматическую идентификацию и удаление дублей публикаций и лишних символов в файле. Для формирования массива ключевых слов программа берет информацию из полей “DE” - Keywords, “ID” - Keywords Plus, а также из названий статей и аннотаций - полей “ТГ” - Title и “AB” - Abstract), что решает проблему неполного покрытия некоторых из этих полей (обозначенную программой Biblioshiny). Программа проводит нормализацию и приведение к единому виду ключевых слов, используя словари для английского языка.

Рабочее окно программы WoS2Pajek

Figure 14: Рабочее окно программы WoS2Pajek

Поскольку программа WoS2Pajek ориентирована на использование информации о цитировании работ (кратких описаний цитируемых публикаций в поле “CR”), фокус делается на обработке библиографических описаний. Работы, указанные в поле “CR”, записаны в формате ISI, предложенном Институтом научной информации (The Institute for Scientific Information), Информация заносится в формате: AU + ', ' + PY + ', ' + SO[:20] + ', ' + VL + ', ' + BP (автор, год публикации, до 20 символов источника публикации / журнала, выпуск, начальная страница) - например, TOSHCHENKO ZT, 2000, SOTSIOL ISSLED, V23, P123. Изначально такой подход использовался для повышения точности данных при внесении информации по единому формату. Но т.к. по факту одна работа может иметь отличающиеся ISI-наименования, для повышения точности программа WoS2Pajek использует короткие имена, записываемые в формате: LastNm[:8] + ' ' + FirstNm[0] + ' (' + PY + ')' + VL + ':' + BP (8 символов фамилии, первая буква имени, год публикации, выпуск журнала, начальная страница) - например, TOSHCHEN\_Z(2000)23:123. Та же самая процедура создания коротких имен осуществляется и для работ, имеющих полные библиографические описания (“хитов”). Для решения проблемы дизамбигуации имен авторов записываются по форме: LastNm[:8] + ' ' + FirstNm[0] (8 символов фамилии, первая буква имени) - например, TOSHCHEN\_Z. Безусловно, при таком подходе могут возникать проблемы “склейки” имен авторов, однако эти проблемы разрешаются путем проверки результатов, получаемых для наиболее важных единиц анализа<sup>7</sup>. Чистка данных, как правило, осуществляется итеративно – при нахождении проблем правки либо вносятся в исходный файл, который

<sup>7</sup>Методология следует т.н. статистическому подходу, согласно которому даже при некоторой неконсистентности в данных общие тренды и важные единицы анализа могут проявиться при анализе.

снова проходит через программу WoS2Pajek, либо устраняются алгоритмическим образом в программе Pajek.

Предварительный анализ массива данных РИНЦ показал, что некоторые важные для анализа параметры метаданных имеют отсутствующие значения (0 или “none”): из 37,302 уникальных авторов в начальном массиве у 17,201 авторов не имелось РИНЦ ID (хотя наличие именно этой информации рассматривалось как преимущество базы). В связи с этим возникла необходимость проведения предобработки данных перед дальнейшим анализом. В ходе преобразований с целью дезамбигуации имен авторов собранная база трансформировалась: вначале была проведена нормализация имен авторов и их аффилиаций, а затем созданы универсальные ID для авторов в формате: eLibrary\_ID + FirstNm[:2] + LastNm[:8] + Affiliation\_ID (ID автора в РИНЦ, инициалы, 8 символов фамилии, ID организации автора) - например, 1382\_ZHT\_Toschenk\_5350 (подробнее см. [467]). В результате предобработки количество уникальных названий аффилиаций сократилось на 39,5% за счёт нормализации и приведения к единому виду описаний аффилиаций и создания универсальных описаний для аффилиаций с единым ID; количество уникальных ID организаций сократилось на 1,5% – за счёт удаления некорректно заполненных ID, а количество уникальных авторов увеличилось на 95% – с 19,366 до 37,790 – за счет идентификации авторов, не имеющих РИНЦ ID.

**Построение сетевых файлов.** Подход к работе с данными WoS CC подразумевает использование программы WoS2Pajek [40] для трансформации массива в коллекцию связанных сетей: двумодальных сетей «Работа – Автор» **WA**, «Работа – Ключевое слово» **WK**, «Работа – Журнал» **WJ** (где в первом наборе указаны все публикации, во втором – авторы, ключевые слова или журналы, а далее фиксируются связи между ними) и одномодальной сети цитирований между работами **Cite**. Также создаются файл с информацией о годах публикации работ (Year.clu), на основании которого сети можно разделить на периоды для изучения в динамике, и файл с разделением работ на источники с полным библиографическим описанием (“хиты”) и только цитируемую литературу (DC.clu), что необходимо для изучения вопросов, связанных с цитированием<sup>8</sup>.

По этой же логике, после формирования массива в РИНЦ из соответствующих полей с помощью специально написанного программного кода в Python были выгружены данные для построения двумодальных сетей «Работа – Автор» **WA** и «Работа – Ключевое слово» **WK**. Файлы сохранены в формате .net и доступны для дальнейшего анализа в программе Pajek. Был сформирован файл с информацией о годах публикации отобранных работ<sup>9</sup>.

Ввиду автоматизированности процесса, процедура предобработки данных и построения сетей с помощью программы WoS2Pajek занимает гораздо меньше времени (так, построение сетевых файлов из начального подмассива заняло 2 мин. 34 сек). Подход к предобработке и подготовке сетевых файлов из массива eLibrary занял гораздо больше времени. Вместе с тем использование разработанного в рамках проекта программного кода может значительно ускорить процесс работы с данными.

<sup>8</sup>Еще раз подчеркнем, что такая информация не приводится в описании РИНЦ, однако теоретически может быть получена путем перехода по имеющимся в записях ссылках на публикации на eLibrary, сбора данных об используемых источниках с помощью специально написанного парсера и последующей чистке массива.

<sup>9</sup>Файлы создавались для проводимого анализа, но на основе имеющейся информации могут быть построены и другие двумодальные сети и дополнительные файлы с атрибутами узлов (количество страниц, цитирование), которые можно использовать для решения разных исследовательских вопросов.

## 6.3 Результаты исследования

### 6.3.0.1 Сравнительный Анализ массивов данных Полнота библиографических метаданных.

Безусловно, не все библиографические описания в базах WoS и eLibrary содержат все возможные метаданные, обозначенные в Табл. ???. Вместе с тем полнота представления метаданных библиографических описаний оказывает значительное влияние на качество анализа, поэтому важно оценить объем пропущенных данных в массивах. Табл. ?? предstawляет данные для оценки полноты библиометрических описаний публикаций в двух рассматриваемых массивах - количество и долю пропущенных значений по отобранным метаданным<sup>10</sup>. Для удобства метаданные сгруппированы по типам библиометрических единиц анализа, к которым они относятся, как в Табл. ???. Обратим внимание, что данные подсчитаны для публикаций: отсутствующие значения по именам авторов показывают количество статей, в которых не имеется хотя бы одного имени автора на русском и английском языках; данные в названиях аффилиаций подсчитывают число случаев, когда при наличии авторов хотя бы у одного из них отсутствует название аффилиации<sup>11</sup>.

Согласно оценке программы Biblioshiny, доли пропущенных значений от 20% до 50% говорят о слабой, а более 50% - о критической представленности данных в поле библиографического описания, и не рекомендуются программой для использования в анализе. Как видим, для массива WoS это (от наибольшей доли пропущенных значений к наименьшей) информация о финансировании, дополнительным ключевым словам, DOI (критично), ключевым словами и аннотации публикации на английском языке (слабо). Для массива eLibrary это информация о финансировании, названию на английском языке, DOI, названию аффилиации автора на русском (критично), аннотации на русском и английском, ключевым словам на английском языке (слабо). Ключевые слова на русском языке отсутствуют в 15% публикаций, однако этот показатель считается “проходным”.

Table 7: Оценка полноты библиографических метаданных в массивах WoS и РИНЦ: пропущенные значения

Метаданные / Пропущенные значения	WoS		РИНЦ	
	Число	%	Число	%
<b>Публикация</b>				
Название - английский язык	0	0	67,048	89.1
Название - русский язык	-	-	0	0
Аннотация - английский язык	737	20.71	27,739	36.9
Аннотация - русский язык	-	-	27,751	36.9
Ключевые слова - английский язык	774	21.75	27,600	36.7
Ключевые слова - русский язык	-	-	11,568	15.4
Дополнительные ключевые слова (поле Keywords Plus)	2806	78.84	-	-
DOI	1780	50.01	62,247	82.7
Год публикации	0	0	0	0
Количество цитирований работы	0	0	0	0

<sup>10</sup>За основу структуры взята таблица, формируемая в программе Biblioshiny при загрузке массива данных. Однако также программа позволяет выгрузить массив в формате Excel, что позволяет самостоятельно оценить заполненность большего количества полей (что было сделано в данном случае и приведено в таблице).

<sup>11</sup>Если автора нет (none и по столбцу с английской фамилией, и по столбцу с русской), наличие аффилиации не проверяется.

Количество цитируемых статей	0	0	-	-
Цитируемые источники	235	6.60	-	-
Число страниц	0	0	0	0
Информация о финансировании	2951	82.9	73,447	97.6
Автор(ы)				
Автор (фамилия и имя) - английский язык	0	0	35	0.05
Автор (фамилия и имя) - русский язык	-	-	1,539	2.05
Название аффилиации автора на русском	-	-	4,609	6.1
Название аффилиации автора на английском	2	0.06	37,891	50.4
Журнал				
Журнал (название)	0	0	0	0
Журнал (ID)	-	-	0	0
Информация об издателе	0	0	1,223	1.6%

*Примечание: Знак “-” означает, что данное поле в базе не представлено. В ключевых словах на английском языке для WoS указаны данные из поля “DE”. Более темным градиентом выделены более высокие доли пропущенных данных.*

**Основная информация по массивам.** В Табл. ?? собрана основная информация по числу различных единиц анализа в рассматриваемых массивах. Обратим внимание, что для сравнения в eLibrary взяты данные по числу ключевых слов на английском языке (аналогичный показатель на русском языке составляет 100,594); ключевые слова были взяты в формате, приведенном авторами, и не подвергались обработке (чем можно объяснить их большое количество, из-за наличия множества уникальных слов). Также приводится два варианта подсчета данных по массиву WoS, в связи с использованием для работы с данными двух программ – Biblioshiny и WoS2Pajek. Если выделенное этим программами число публикаций и журналов является одинаковым, то по количеству выделенных авторов и ключевых слов наблюдаются различия, связанные с алгоритмами предобработки, имплементированным в программы. Так, у Biblioshiny количество авторов составляет 3,554, а у WoS2Pajek – 3,238<sup>12</sup>, а количество ключевых слов – 12,215 и 6,750, соответственно. Если первые различия являются незначительными, то во втором случае данные, полученные WoS2Pajek, следует рассматривать как более валидные, ввиду более точного подсчета (для Biblioshiny число рассчитано как сумма по полям ID и DE, нет возможности учета пересечений) и заложенных в программу алгоритмов нормализации ключевых слов. Для информации подсчитано соотношение единиц из массива WoS к массиву eLibrary, показывающее значительное превосходство данных eLibrary по объему и небольшое число потенциально пересекающихся сущностей.

Table 8: Число единиц анализа в массивах WoS и eLibrary

Единица анализа / Количество	eLibrary	Biblioshiny	WoS2Pajek

<sup>12</sup>Изначально – 3,241 автора; при эксплораторном анализе в Pajek было идентифицировано несколько проблем и имена авторов были почищены вручную.

WoS	% WoS к eLibrary	WoS	% WoS к eLibrary		
Публикации	75,232	<b>3,559</b>	4.7	<b>3,559</b>	4.7
Журналы	3,910	<b>109</b>	2.8	<b>109</b>	2.8
Авторы	37,790	3,554	9.4	3,238	8.6
Ключевые слова на английском	91,109	12,215	13.4	6750	7.4

Динамика количества **публикаций** в обеих базах за рассматриваемый период показана на Рис.?? - ??.

Распределение абсолютного числа публикаций (Рис. ??) в eLibrary показывает плавный рост и достижение максимума в 2016 г. и следующее за ним снижение. Количество российских социологических публикаций в WoS достигает максимума в 2019 г., однако далее снижается незначительно. Чтобы увидеть общие тренды, данные были подсчитаны кумулятивно и затем нормированы в диапазоне от 0 до 1 (значение за каждый год разделено на сумму публикаций). В такой репрезентации лучше видно, что относительные доли числа публикаций в eLibrary с 2014-2015 гг. были выше, чем в WoS. Однако если средний годовой прирост публикаций<sup>13</sup> в eLibrary на 2021 г. составляет -0.6%, аналогичный показатель для WoS составляет 5.5%, что говорит о более динамичном увеличении числа публикаций в этой базе.

Динамика количества публикаций в базах WoS и eLibrary: абсолютное число публикаций (3-а, сверху) и нормированное на кумулятивной шкале (3-б, снизу)313\_5) Динамика количества публикаций в базах WoS и eLibrary: абсолютное число публикаций (3-а, сверху) и нормированное на кумулятивной шкале (3-б, снизу)313\_4)

Анализ пересечений между базами на основе статистики

**Анализ баз данных, индексированных в eLibrary.** Интересной находкой исследования стало то, что для каждой публикации в массиве eLibrary содержится информация о том, в каких научометрических базах проиндексирован опубликовавший ее журнал, что делает возможным посмотреть на распределение и пересечение баз на одной площадке. Как видно из Табл. ??, большинство публикаций опубликованы в журналах, индексируемых в базе РИНЦ (85%) и входящих в список ВАК (53%). Значительно меньшее количество статей опубликованы в журналах, индексируемых в RSCI (11%), Scopus (8%) и WoS CC (7%). Множества, составляемые массивами публикаций в разных базах, являются пересекающимися. На основе имеющихся данных можно посмотреть на пересечение (как множество общих единиц) и объединение (как множество всех единиц) между различными базами на уровне публикаций и журналов.

Table 9: Индексация публикаций из массива eLibrary в различных базах {#tbl:indexElib}

База	Показатель	Входит в базу	Не входит в базу	Сумма

<sup>13</sup>Рассчитанный путем деления разницы между количеством публикаций за каждую пару лет на начальное значение, и расчет среднего за анализируемый период времени.

<b>РИНЦ / RISC</b>	абсолютные значения	65,103	10,129	75,232
%	87%	13%	100%	
<b>ВАК</b>	абсолютные значения	40,046	35,186	75,232
%	53%	47%	100%	
<b>RSCI</b>	абсолютные значения	8,179	67,053	75,232
%	11%	89%	100%	
<b>Scopus</b>	абсолютные значения	6,222	69,010	75,232
%	8%	92%	100%	
<b>WoS</b>	абсолютные значения	5,226	70,006	75,232
%	7%	93%	100%	

*Сходство баз на уровне публикаций.* Разные комбинации баз данных составляют разные доли от общего числа статей в массиве eLibrary (Табл. {#tbl:simElib}). Некоторые включенности одних множеств в другие объясняются известной информацией о создании баз: так, все публикации из RSCI по дефолту включены в базу РИНЦ, поскольку являются подмножеством статей, опубликованных в российских топ-журналах. Известно, что база РИНЦ включает публикации российских авторов, опубликованные в журналах WoS и Scopus – публикации в этих базах также полностью (5,226 в WoS) и почти полностью (6,212 из 6,222 в Scopus) входят в РИНЦ. Ситуация для базы RSCI несколько иная: из 8,179 публикаций в этой базе российских топ-журналов в журналах WoS CC также индексированы 4,263 (52%) работ, а из 5,226 публикаций 963 статьи (18.4%) не входят в RSCI, а индексируются только в WoS CC (однако попадают в базу благодаря тому, что их индексирует РИНЦ). Число статей из базы RSCI, также индексированных в базе Scopus, составляет 5,249 (64.2% от всех публикаций в RSCI), а число уникальных статей из Scopus в нашей базе составляет 973 статьи (15.6% от всех публикаций в Scopus). Общее пересечение статей, входящих в WoS, и в Scopus, составляет 4,170 статей – 79.8% от всех публикаций WoS в массиве, и 67.0% от всех публикаций в Scopus. Аналогичная доля рассчитывается и на пересечении этих двух баз и РИНЦ (опять же, по природе создания базы), но если сравнивать с базой RSCI, то число общих статей на пересечении трех баз составляет 3,875 (что составляет 74.1% от всех статей в WoS, 62.3% – в Scopus, и 47.4% – в RSCI). Общее же число публикаций, представленных во всех трех базах (RSCI, Scopus, WoS), которые составляют ядро РИНЦ, рассчитанное как объединение множеств, составляет 9,820 публикаций - 13% от всех публикаций в массиве.

Обращает на себя внимание интересный факт: Табл. {#tbl:indexElib} показывает, что не все статьи, вошедшие в массив eLibrary, входят в базу РИНЦ (только 87%). Предполагая, что оставшиеся 13% статей распределены по другим базам, указанным в массиве eLibrary, мы посмотрели на объединения баз Scopus, WoS и РИНЦ, а также объединение всех пяти баз (Табл. ??). Выяснилось, что первое объединение составляет 65,113 публикаций, а второе – 65,308 публикаций - то есть снова около 87%

публикаций из базы. Оставшиеся 9,924 статей не входят ни в одну из пяти баз, указанных на eLibrary. Более внимательный анализ этого подмассива публикаций показал, что они опубликованы в журналах, которыми заключено лицензионное соглашение на размещение издания на eLibrary.ru. Кроме того, выборочный анализ некоторых журналов с помощью системы SCIENCE INDEX на eLibrary показал, что в определенные периоды времени эти журналы индексировались в РИНЦ. Топ журналов из данного подмассива приведен в Табл. {#tbl:journElib}; ведущим источником выступает “Экономика и социум” с 1759 статьями. Таким образом, в ходе анализа была уточнена реализованная стратегия сбора данных: отбор статей, индексируемых eLibrary, не аналогичен отбору по статьям, индексируемым в РИНЦ.

Наконец, еще одно пересечение баз данных относится к публикациям, входящим в список ВАК. Всего в массиве eLibrary 40,046 публикаций из этой базы, и их подавляющее большинство (99.5%) входят в РИНЦ; разница между базами составляет 205 журналов. При объединении же множества ВАК со Scopus, WoS и RSCI получается 40,292 публикации - всего на 246 больше, чем в базе ВАК; получается, что список ВАК в значительной степени состоит из журналов, индексируемых в этих трех базах (ядре РИНЦ). Результат объединения пяти баз в числовом выражении аналогичен объединению множеств РИНЦ и ВАК – то есть все статьи, индексируемые в ядре РИНЦ, сюда входят.

Table 10: Сходство между базами данных по числу публикаций (массив eLibrary) {#tbl:simElib} \*\*\*\*\*

База	Число публикаций	% от общего числа стат.
<b>Пересечение (множество общих статей - правило “И”)</b>		
РИНЦ + RSCI	<b>8,179</b>	10.9%
РИНЦ + WoS	<b>5,226</b>	6.9%
РИНЦ + Scopus	<b>6,212</b>	8.3%
RSCI + WoS	<b>4,263</b>	5.7%
RSCI + Scopus	<b>5,249</b>	6.98%
WoS + Scopus	<b>4,170</b>	5.5%
Scopus + WoS + РИНЦ	4,170	5.5%
Scopus + WoS + RSCI	3,875	5.2%
РИНЦ + ВАК	39,841	52.6%
<b>Объединение (множество всех статей - правило “ИЛИ”)</b>		
Scopus + WOS + RSCI (ядро РИНЦ)	9,820	13%
Scopus + WOS + РИНЦ	65,113	86.5%
Scopus + WOS + RSCI + ВАК	40,292	53.6%
Scopus + WOS + RSCI + ВАК + РИНЦ	65,308	86.8%
РИНЦ + ВАК	65,308	86.8%

Table 11: Топ журналов из подмассива публикаций, индексируемых только в eLibrary

№	Название журнала	Количество статей	№	Название журнала	Количество статей

1	Экономика и социум	1759	12	Студенческий	166
2	Молодой ученый	441	13	Гуманитарные научные исследования	145
3	Сборники конференций НИЦ Социосфера	425	14	Современные тенденции развития науки и технологий	137
4	Аллея науки	327	15	Студенческий вестник	122
5	NovaInfo.Ru	250	16	Научный альманах	119
6	Вестник современных исследований	193	17	Вестник научных конференций	116
7	Теория и практика современной науки	176	18	Евразийский союз ученых	115
8	Актуальные проблемы гуманитарных и естественных наук	176	19	Современные научные исследования и инновации	110
9	Стратегия устойчивого развития регионов России	174	20	Форум молодых ученых	106
10	Система ценностей современного общества	171	21	Colloquium-journal	106
11	Сборник научных трудов SWorld	168	22	Альманах современной науки и образования	99

*Сходство баз на уровне журналов.* Количество журналов, индексируемых в разных базах по массиву eLibrary, показано в Табл. ??.<sup>14</sup> Подавляющее число источников (91.6%), в которых опубликованы статьи в базе, индексируются в РИНЦ; доли журналов, индексируемых в базах RSCI, Scopus, WoS достаточно небольшие и составляют 4% - 5% от всех журналов. Треть всех журналов включены в список ВАК.

По анализируемому массиву данных из всех журналов в РИНЦ, где опубликованы работы по социологии, в число топ-журналов, отобранных для базы RSCI, входит 202 журнала. Число журналов из РИНЦ, индексируемых в WoS, составляет 148 журналов, а в Scopus - 193 журнала; на пересечении эти две зарубежные базы дают в РИНЦ 92 журнала (что составляет 48% от всех журналов в Scopus и 62% от всех журналов в WoS). Число журналов из RSCI на пересечении с WoS дает 51 журнал, а со Scopus - 83, на пересечении трех баз находится 41 журнал.

Table 12: Индексация журналов из массива eLibrary в различных базах

<b>База</b>	<b>Количество журналов</b>	<b>% от общего числа журналов</b>
<b>РИНЦ</b>	<b>3580</b>	<b>91.56%</b>
<b>RSCI</b>	<b>202</b>	<b>5.17%</b>
Scopus	193	4.94%
<b>WoS</b>	<b>148</b>	<b>3.79%</b>
BAK	1310	33.5%
<b>Всего журналов</b>	<b>3910</b>	<b>100%</b>

Полученные результаты демонстрируют, что база РИНЦ, максимально близкая по размеру базе eLibrary (но не полностью покрывающая ее), в значительной степени пересекается с другими, меньшими по размеру базами библиографических данных, в т. ч. базой WoS. Далее проверим наличие такого пересечения путем сравнения двух рассматриваемых в работе массивов.

**Анализ массивов данных WoS и eLibrary.** Сравнение массивов проводится по публикациям и авторам, включенными в два анализируемых массива данных.

*Сопоставление публикаций.* Для сопоставления публикаций, входящих в базы WoS (3,559) и eLibrary (75,232), было использовано несколько подходов: мы сопоставляли данные по: 1) DOI; 2) названию публикаций на английском языке; 3) генерированной комбинации из последовательности авторов и года написания статьи.

Поиск совпадающих DOI статей позволил сопоставить 655 публикаций. Ситуацию осложняло отсутствие DOI у 50% статей в WoS и у 83% из eLibrary. Далее сопоставление статей было продолжено путем сравнения их названий на английском (обратим внимание на высокую долю пропущенных значений). Названия были предварительно предобработаны (убраны знаки препинания и цифры, все символы приведены к нижнему регистру, убраны стоп-слова: the, of, in, at, is; артикли), и по точным совпадениям удалось получить еще 32 совпавшие статьи. Далее мы приступили к поиску совпадений по комбинации авторов и года написания статьи. Например, если Михаил Соколов (*SOKOLOV MM*) и Кирилл Титаев (*TITAEV KD*) написали статью в 2014 году, их статья была присвоена строка “*SOKOLOV MM; TITAEV KD\_2014*”. Такие “простые ID” мы генерировали для всех статей в WoS и eLibrary, и искали

<sup>14</sup>Подсчет количества журналов осуществлялся по полям с уникальными ID журнала.

совпадения по ним. Отметим, что для повышения точности оценки и избежания ложно-положительных совпадений, поиск совпадений проводился только для статей, чье simple\_ID встречалось в базе данных только один раз. В противном случае совпадение будет неточным: один и тот же социолог или группа исследователей может написать несколько статей в один и тот же год, и мы не сможем точно сопоставить, например, 1 публикацию Ж. Т.Тощенко в 2012 году из WoS с какой-то из 6 публикаций Тощенко в 2012 году в eLibrary. По результатам этого поиска нашлось ещё 358 статей.

Несмотря на то, что комбинация выбранных способов поиска идентичных статей неидеальна, она позволяет получить примерную оценку совпадения двух баз данных без разработки специфических технологических решений. В теории они могли бы включать поиск совпадающих статей по разным вариациям автоматического перевода названия статьи с русского на английский, поиска совпадающих статей по комбинации авторов с поиском возможных расхождений в 1-2 символа в фамилиях, и пр., однако эта разработка может стать темой отдельного исследовательского проекта. Итоговое значение совпавших статей в базе данных elibrary и WoS составляет 1013 статей – 28.5% от всех публикаций в базе данных WoS или 1.4% от всех публикаций в базе данных elibrary.

*Сопоставление авторов.* Для сравнения авторов, присутствующих в базах WoS и eLibrary, мы выбрали подход, в котором искали совпадения по фамилиям и инициалам авторов; таким образом, ограничением для следующих оценок стало предположение о том, что одна комбинация фамилии и инициалов принадлежит одному автору. Так, хотя в предварительно обработанных данных eLibrary были созданы новые универсальные ID, позволяющие идентифицировать разных авторов (основываясь на ID РИНЦ, инициалах, фамилии и ID аффилиации), они бы не совпадали с потенциальными ID, которые можно было бы сконструировать на основе данных WoS. WoS содержит информацию о ResearcherID и ORCID-ID исследователей, но эти поля часто не заполнены. В нашей базе в 54.5% статей нет информации о ResearcherID ни для одного из авторов статьи; для ORCID-ID аналогичная оценка составляет 61.4%.

По этим причинам в базе данных статей российских социологов из eLibrary мы создали столбец с перечислением всех авторов, аналогичный по формату записям в базе WoS, где авторы записываются в столбце “AU” следующим образом: “KOLESNIK NV;SHOPULATOV AN;SINYUTIN MV”. Затем был проведен поиск совпадений по авторам, чтобы узнать, какие авторы присутствуют только в одной или обеих базах. Отметим, что предобработка фамилий (например, приведение отдельных написаний фамилии к наиболее популярному виду) не производилась, однако такие процедуры можно было бы провести, тем самым повысив точность оценки. По полученным результатам (Табл. ??), число авторов в обеих базах составило 1,180 авторов, что составляет 33% от всех авторов в массиве WoS, но только 3.3% от всех авторов.

Table 13: Показатели сопоставления авторов в базах WoS и eLibrary {#tbl: auWosElib}

Показатель	Значение
Число авторов в eLibrary	35 462
Число авторов в WoS	3 554
Число авторов в обеих базах	1 180
Доля авторов в обеих базах относительно числа уникальных авторов в данных WoS	33%

Доля авторов в обеих базах относительно числа уникальных авторов в данных eLibrary 3.3%

*Примечание: Число авторов в eLibrary подсчитано по аналогии подсчета для WoS*

**6.3.0.2 Анализ Пересечений между базами на основе содержательных результатов** В данном разделе проводится опосредованное, непрямое сравнение того, насколько похожими являются массивы данных WoS и eLibrary с точки зрения получаемых результатов при анализе массива и производных сетей.

**Работы и авторы.** Входящая центральность в двумодальной сети **WA** показывают количество работ у авторов. Распределение этого показателя для двух массивов приведено на Рис. 15. Массивы данных значительно различаются по размеру – число публикаций в массиве eLibrary примерно в 20 раз больше числа публикаций в WoS – поэтому различия наблюдаются и в числе авторов. Вместе с тем распределения на Рис. 15 похожи по тренду и могут следовать степенному закону, или закону Лотки, описывающему распределение продуктивности ученых<sup>15</sup>. Тогда как 66% авторов в массиве eLibrary и 64% в массиве WoS имеют только одну публикацию, еще 13% и 14% - две, а по 6% - три, некоторые авторы в базах являются супер-продуктивными, имея 241, 2015 и 192 публикации в базе eLibrary и 45, 29 и 28 публикаций в базе WoS. Наиболее продуктивные авторы с наибольшим количеством работ по двум массивам приведены в Табл. {#tbl:aumost}. Четверо из выделенных топ-20 авторов присутствуют в обеих базах, однако авторы из массива eLibrary, имеющие наибольшее количество публикаций, в базе WoS имеют 1, 1 и 5 публикаций.

Распределение количества работ по авторам в двух массивах данных (логарифмическая шкала)

Figure 15: Распределение количества работ по авторам в двух массивах данных (логарифмическая шкала)

Table 14: Авторы с наибольшим количеством публикаций, по двум массивам {#tbl:aumost}

Rank	Id	Value	Rank	Id	Value
1	TROTSUK_I	45	1	429210_SI_Samygin_14461	241
2	PUZANOVA_Z	29	2	74486 SG_Maksimov_258_7082	215
3	KRAVCHEN_S	28	3	767943_TK_Rostovsk_924_1432_1488_4812_5350_13701	192
4	TOSHCHEN_Z	20	4	<b>137655 GE_Zborovsk_290_1255_7366_14141</b>	166
5	NARBUT_N	18	5	75266_NV_Dulina_306_1000	160
6	ZBOROVSK_G	18	6	145046_OE_Nojanzin_258_7082	150
7	SOROKIN_P	18	7	72232_JUG_Volkov_322_1432_3455_14829	147
8	SOKOLOV_M	18	8	129623_VA_Il'in_815	142
9	GORSHKOV_M	17	9	287431_AV_Verescha_322_1432_14461	133
10	YANITSKI_O	17	10	504328_MV_Morev_815	120
11	ROMANOVS_N	16	11	<b>251886 IV_Trotsuk_421_425</b>	120

<sup>15</sup>Согласно закону Лотки, число авторов, опубликовавших в течение определенного периода  $n$  статей, обратно пропорционально квадрату  $n$ . Этот закон можно проверить математической функцией.

12	TESLYA_A	15	12	495445_DA_Omel'che_258	119
13	KOZYREVA_P	14	13	<b>1382_ZHT_Toschenk_5_5350</b>	117
14	SMIRNOV_A	14	14	73979_HV_Dzutsev_1432_4812	116
15	OBRAZTSO_I	14	15	442046_JUV_Stavropo_259_808	116
16	TIKHONOV_N	13	16	674856_NH_Gafiatul_322_761	111
17	GASPARIS_A	13	17	265785_VP_Babintse_340_1279_6227	110
18	RYBAKOVS_L	13	18	331427_SA_Il'inyh_1068	109
19	LAPIN_N	13	19	<b>72610_MK_Gorshkov_1432_14554</b>	109
20	LARINA_T	13	20	259120_PA_Ambarova_290	106

Показатель исходящей центральности в сети **WA** показывает количество авторов в работах (Табл. ??). Максимальное число авторов в массиве WoS составляет 15; для массива eLibrary при сборе данных было установлено ограничение в 8 авторов. Как видно, доли публикаций статей с единственным автором в двух массивах являются практически идентичными - 62% в WoS и 66% в eLibrary. Это подтверждает обозначенную гипотезу о распространенности практики публикаций с единственным автором как части публикационной культуры в области социальных наук. Следующий самый часто встречающийся в публикациях формат - подготовка публикаций парами авторов, - встречается в 24% и 25% статей в WoS и eLibrary, соответственно; за ним следуют публикации, сделанные тремя (9% для WoS и 7% для eLibrary) и четырьмя (3% и 1.5%) авторами. Статьи с относительно большим количеством авторов встречаются в массивах в единичном виде.

Table 15: Количество авторов в публикациях в двух массивах {#tbl:auamount}

WoS			eLibrary		
Число авторов	Частота	Доля, %	Число авторов	Частота	Доля, %
1	2217	62.29	1	49973	66.43
2	844	23.71	2	18473	24.55
3	327	9.19	3	5044	6.7
4	120	3.37	4	1144	1.52
5	34	0.96	5	361	0.48
6	5	0.14	6	113	0.15
7	6	0.17	7	63	0.08
8	2	0.06	8	61	0.08
9	1	0.03			
12	1	0.03			
14	1	0.03			
15	1	0.03			

**Коллaborации авторов.** На основе сети **WA** путем ее перемножения может быть построена базовая ненормализованная сеть соавторства **Co**, где сила связей рассчитывается исходя из количества

публикаций, рассчитываемых авторами совместно, а петля обозначает общее количество работы у авторов, написанных в соавторстве и самостоятельно [9]. Доли авторов, не имеющих хотя бы одного соавтора, для массивов eLibrary и WoS составляют соответственно 35.8% и 27.8% (т.е. для массива WoS доля чуть ниже). Распределение по числу соавторов у авторов (Рис. 16) показывает, что большинство из них имеют одного (31% в eLibrary и 27% в WoS), двух (14% и 18.5%) или трех соавторов (6.5% и 12.5%). Однако выделяются авторы со значительным количеством соавторов: в массиве WoS - Н.Е. Покровский (25 соавторов), В.В. Щербина (21) и Ж.Т. Тощенко, Н.В. Романовский и А.Б. Гофман (20), в массиве eLibrary доля авторов с числом соавторов более 20 составляет 0.57%, или 212 авторов, среди которых лидирует С.И. Самыгин со 140 соавторами.

Распределение количества соавторов по авторам в двух массивах данных (логарифмическая шкала)

Figure 16: Распределение количества соавторов по авторам в двух массивах данных (логарифмическая шкала)

**Работы и журналы.** В Табл. ?? приведены топ-25 журналов по количеству публикаций в двух массивах. Лидером в обеих базах выступает журнал “Социологические исследования” - абсолютное количество публикаций в нем в WoS и eLibrary примерно одинаково (1,923 и 1,905, соответственно). Если же посмотреть на вклад журнала в общее количество публикаций, то значение этого источника для базы WoS становится еще важнее – публикации в нем составляют 54% от всего массива данных. В eLibrary вклад “Социса” растворяется в связи с большим количеством журналов; несколько других журналов с большим вкладом идут с довольно небольшим отставанием. На основе распределения журналов из массива WoS видно, что вклад российских авторов на эту площадку (зарубежную “витрину”) в основном делается через публикации в российских журналах, индексируемых в WoS (первые пять российских журналов составляют 90% публикаций).

Table 16: Топ-15 журналов в двух массивах данных по числу публикаций {#tbl:topjourn}

Ранг	WoS			eLibrary		
	Журнал	N	% от всех публикаций	Журнал	N	% от всех публикаций
1	SOTSIOL	1923	54.0%	Социологические исследования	1905	2.5%
	ISSLED+					
2	RUDN J	546	15.3%	Экономика и социум	1759	2.3%
	SOCIOL					
3	SOCIOL	309	8.7%	Теория и практика общественного развития	1078	1.4%
	OBOZR					

4	J ECON SOCIOL	245	6.9%	Гуманитарные, социально-экономические и общественные науки	911	1.2%
5	SOCIOL NAUK TEHNOL	167	4.7%	Социально-гуманитарные знания	863	1.1%
6	CHANG SOC PERSONAL	51	1.4%	Социология	737	1.0%
7	INT J SOCIOL SOC POL	37	1.0%	Мониторинг общественного мнения: экономические и социальные перемены	731	1.0%
8	COMP SOCIOL	23	0.6%	Социология в современном мире: наука, образование, творчество	670	0.9%
9	INT J IN- TERCULT REL	22	0.6%	Социальная политика и социология	630	0.8%
10	SOC INDIC RES	20	0.6%	Социальные и гуманитарные науки.	598	0.8%
11	FILOS- SOCIOL	18	0.5%	Отечественная и зарубежная литература. Серия 11: Социология Юга России	582	0.8%

12	SPORT SOC	8	0.2%	Власть	528	0.7%
13	POETICS	8	0.2%	Журнал социологии и социальной антропологии	523	0.7%
14	CORVINUS J SOCIO PO	5	0.1%	Общество: социология, психология, педагогика	494	0.7%
15	CURR SOCIOL	5	0.1%	Известия Саратовского университета. Новая серия. Серия: Социология. Политология	482	0.6%

---

**Работы и ключевые слова.** Показатель исходящей центральности в сети **WK** показывает количество ключевых слов в работе. Для работ из массива WoS этот показатель варьируется от 1 до 40, а из массива eLibrary - от 1 до 51 (при этом в 36.7% случаев значения пропущены). Показатель входящей центральности в сети **WK** показывает частоту использования различных ключевых слов в работах. Как показывает распределение этих значений для двух массивов (Рис. 17), 77% ключевых слов в массиве eLibrary и 50.5% в WoS использованы только один раз, еще 10% и 14% соответственно - два раза, 3.6% и 7% - три раза, и т.д. В Табл. ?? приведены топ-20 слов, наиболее часто используемых в обоих массивах с долей к числу всех ключевых слов. Повторяющиеся слова из двух массивов выделены цветом.

Частота использования ключевых слов в работах в двух массивах данных (логарифмическая шкала)

Figure 17: Частота использования ключевых слов в работах в двух массивах данных (логарифмическая шкала)

Table 17: Топ-20 ключевых слов для обоих массивов {#tbl:topkeyw} \*\*\*\* Примечание: выделены **полностью** или частично повторяющиеся слова. ### Обсуждение и выводы Обзор исследований по сравнению баз данных научных публикаций подтверждает, что даже при наличии альтернатив WoS является одним из самых популярных источников информации для наукометрических исследований. Безусловными плюсами работы с этой базой является наличие инструментов для выгрузки, предобработки и статистического и сетевого анализа публикаций, но, в случае с данными российских авторов, минусом – ограниченная представленность публикаций. В eLibrary, напротив, отечественные публикации представлены максимально полно (и не ограничиваются только публикациями в научных журналах и главами в монографиях); проблема заключается в отсутствии широко доступных сервисов по обработке и анализу данных для этой базы. В этой ситуации у исследователя, нацеленного на изучение современного состояния развития российской науки, возникает ряд вопросов: – Можно ли взять только одну базу в качестве источника информации, или необходимо комбинировать данные из нескольких баз? – В случае использования одной базы, насколько валидными будут полученными результаты? – Если данные должны комбинироваться, то как именно это нужно делать?

Ранг	eLibrary		WoS	
	Слово	Значение	Id	Value
1	youth		2849	social
2	<b>society</b>		1343	russian
3	family		1269	<b>sociology</b>
4	values		1225	<b>russia</b>
5	<b>culture</b>		1035	<b>society</b>
6	<b>education</b>		990	sociological
7	globalization		876	analysis
8	students		855	theory
9	migration		839	study
10	socialization		820	<b>education</b>
11	identity		791	<b>state</b>
12	civil society		716	political
13	modernization		696	research
14	communication		631	development
15	<b>state</b>		611	science
16	management		601	life
17	<b>russia</b>		580	<b>cultural</b>
18	region		579	value
19	internet		571	practice
20	<b>sociology</b>		534	public

В нашем исследовании проводится сравнение двух баз через описательный анализ их возможностей по работе с данными и сопоставление двух массивов по одной и той же предметной области – социологии, – что является распространенной практикой в дизайне аналогичных исследований. Полученные массивы данных сравниваются по своей структуре, размеру, полноте метаданных, а также посредством анализа производных базовых сетей. Сравнение результатов анализа с точки зрения содержания позволяет сделать выводы о том, насколько похожие результаты дает использование двух баз данных. Это важно не только с наукометрической точки зрения, но и с позиции изучения ориентаций

ученых на международные и локальные научные сообщества, если думать о двух площадках как о двух возможных направлениях позиционирования ученых.

Несмотря на похожий набор метаданных, базы WoS и eLibrary имеют некоторые различия. Важным преимуществом WoS является наличие списков литературы, что важно для анализа цитирований. В eLibrary авторы и организации имеют ID, однако по факту в большом количестве случаев эта информация отсутствует, что приводит к необходимости предварительной обработки массивов данных. В отличие от данных WoS, работа с которыми может осуществляться в нескольких программах, работа с данными eLibrary как по предобработке, так и по построению сетевых файлов для дальнейшего библиометрического анализа является гораздо более трудозатратой. С точки зрения существующих программ, поскольку в обоих массивах в значимом числе публикаций отсутствует информация о ключевых словах и аннотациях, нужно отметить преимущество, которое, на наш взгляд, дает программа WoS2Pajek по сравнению с BiblioShiny для нормализации ключевых слов, позволяя описать статьи также из их названий (но анализ доступен только для ключевых слов на английском языке).

Массив eLibrary является гораздо более крупным, т. к. в нем содержится информация из разных научометрических баз, и WoS Core Collection является только одной из них, имеющей при этом довольно строгие критерии индексации журналов. Соответственно, число потенциально включенных в базу eLibrary библиометрических сущностей (публикаций, журналов, авторов, ключевых слов) максимально равно количеству этих сущностей в массиве WoS. Анализ массива eLibrary, в котором представлена информация о включенности публикации в разные научометрические базы, показала, что все публикации, индексированные в WoS, входят в РИНЦ, но не все – в RSCI; это говорит о том, что на площадке WoS есть уникальные публикации, которые выступают дополнением к базе RSCI. Поскольку анализ показал, что массив WoS CC более чем на 90% формируется публикациями из российских журналов (которые с высокой вероятностью индексируются в RSCI), непересекающееся множество показывает публикации российских авторов в иностранных индексируемых журналах. Несмотря на то, что по логике рассматриваемые нами для примера базы должны в значительной степени пересекаться (массив WoS должен быть включен в массив eLibrary), пересечение между рассматриваемыми массивами является далеко не полным (около 30% массива WoS входят в массив eLibrary). Это может объясняться как несовершенством реализованной процедуры поиска идентичных публикаций и авторов, так и тем, что в массивах содержатся уникальные публикации и авторы. Эта часть анализа в настоящий момент может рассматриваться как экспериментальная и заслуживает дальнейшей проработки для уточнения пересечения массивов. Отметим, что наличие DOI у всех статей и ResearcherID / ORCID-ID у авторов могло бы существенно упростить эту задачу.

Динамика количества публикаций в двух массивах показывает, что база WoS прирастает более активно. Однако данные в обоих массивах распределяются похожим образом, что говорит о том, что они следуют похожим библиометрическим трендам и законам. Схожим образом в обоих массивах разделяются доли числа работ у авторов (две трети авторов с одной статьей), авторов у работ (две трети работ наблюдаемых в социальных науках «авторов-одиночек», четверть работ, написанных в парах), соавторов у авторов (около трети авторов без соавторов и столько же – с одним соавтором); доля статей с одним ключевым словом в массиве eLibrary выше, чем в WoS (77% против 50%). Выделенные топ-единицы анализа при этом пересекаются только частично, что говорит о наличии своих особенностей

в каждом массиве – самых продуктивных авторов для каждой площадки, наиболее используемых журналов и уникальных ключевых слов, характеризующих исследования. Более подробный анализ наблюдаемых пересечений и отличий может помочь ответить на различные содержательные вопросы о специфике исследований, ориентированных на разные аудитории (хотя возникает вопрос, насколько «ориентированными» на зарубежные исследовательские группы являются публикации в WoS, изначально вышедшие в российских журналах и внесенные в базу благодаря их индексации).

Возвращаясь к вопросу выбора базы данных для анализа нужно сказать, что, как выяснилось в ходе анализа, множество статей в eLibrary не идентично множеству статей РИНЦ, и данные в последнем являются более релевантными. Выбор в качестве основы базы RSCI не включает в массив статьи, опубликованные за пределами российских журналов. Выбор WoS CC в значительной степени ограничивает анализируемый объем статей. В идеальной ситуации анализ публикаций российских авторов должен осуществляться на основе нескольких баз данных – например, комбинация РИНЦ + WoS + Scopus или RSCI + WoS + Scopus (ядро РИНЦ), однако в условиях ограниченных ресурсов выбор может быть сделан в сторону РИНЦ. Если же источником является база WoS, то лучше включать не только Core Collection, но также RSCI и ESCI. Методологические вопросы выгрузки данных, поиска совпадений между базами и их объединения в единый массив пока являются открытыми и требуют дальнейшей разработки.

В качестве общей рекомендации нужно сказать, что исследователь, работающий в области библиометрического анализа, должен хорошо понимать структуру баз, с которыми он работает, чтобы получить нужную ему информацию на входе для дальнейшего анализа, а не просто «искать где светлее» (= в WoS, т. к. разработаны инструменты для анализа), а также ставить исследовательские вопросы с пониманием ограничений в покрытии баз данных. С точки зрения получения валидных результатов важной является также оценка полноты отдельных метаданных в библиографических описаниях.

Дальнейшая работа над этой тематикой требует разработок в области методологии сбора, поиска совпадений и объединения массивов из различных источников. Полученные результаты в области социологии интересно сравнить с другими предметными областями.

## **6.4 Адаптация методов текстового анализа для извлечения информации графовыми методами.**

В современном информационном обществе огромное количество данных порождает необходимость в их эффективной обработке и анализе. Два ключевых инструмента, которые активно применяются для раскрытия информационного потенциала, - это текстовый анализ и графовые методы.

Текстовый анализ, в свою очередь, позволяет извлекать смысл и информацию из текстовых данных, выявлять тенденции, выделять сущности и ключевые аспекты. В контексте современных исследований текстовый анализ становится неотъемлемым инструментом для обработки больших объемов текстовой информации, таких как социальные медиа, новостные статьи, научные публикации и другие.

С другой стороны, графовые методы предоставляют мощный инструмент для визуализации и анализа связей между объектами, для более глубокого понимания контекста. Графовые структуры позволяют представлять информацию в виде узлов и связей между ними, что особенно полезно для

моделирования и анализа сложных систем и взаимодействий. [297]

Например, в статье ‘End-to-end construction of NLP knowledge graph’ [14] авторы описывают методику создания Графа знаний (Knowledge Graph, KG) в области обработки естественного языка на основе научных статей. Основное внимание уделяется извлечению четырех типов связей: “evaluatedOn” между задачами и наборами данных, “evaluatedBy” между задачами и метриками оценки, а также связям “coreferent” и “related” между сущностями одного типа. В статье представлены новые методы для каждого из этих типов связей, а разработанный фреймворк (SciNLP-KG) применяется к 30 000 статьям по NLP из коллекции ACL Anthology для построения графа знаний. Этот граф может автоматизировать создание научных рейтингов для сообщества исследователей в области обработки естественного языка.

Обосновывая важность совместного применения текстового анализа и графовых методов, следует отметить, что современные исследования все больше ориентированы на многомерные данные, представляющие сложные взаимосвязи.

Адаптация методов текстового анализа для работы с графиками становится неотъемлемым инструментом в раскрытии глубокой структуры информации. Это позволяет преобразовать текстовую информацию в структурированные графы, что может обеспечить более глубокий и комплексный анализ, другой взгляд на взаимосвязи и закономерности.

Текстовый анализ включает в себя целый спектр методов для обработки, понимания и извлечения информации из текстовых данных. Одним из ключевых этапов является представление текста в форме, пригодной для структурного анализа. Различные подходы к преобразованию текстовых данных в графовую форму предоставляют эффективные инструменты для визуализации и анализа сложных структур.

Один из подходов к представлению текста в виде графовой модели заключается в трансформации слов или фраз в узлы графа и использовании рёбер для отражения связей между ними. Каждый узел представляет отдельный токен или синтаксическую единицу текста, а рёбра отражают связи и зависимости между этими элементами. Такой графовый подход позволяет сохранить информацию о структуре текста и взаимосвязях между его компонентами, поэтому часто применяется в лингвистических исследованиях для отражения синтаксических отношений в тексте. [293]

Реализация этапа предварительной обработки текста, включающего в себя токенизацию и нормализацию (лемматизацию, удаление шумов и др.), предшествующая созданию графа, предоставляет основу для более эффективного анализа. Токенизация, как процесс, разделяющий текст на токены или элементарные единицы, и лемматизация, направленная на приведение слов к их нормализованным, словарным формам, служат средствами для унификации и стандартизации текстовых данных. Такой подход сокращает размерность графа, устранивая поверхностные вариации слов и подчеркивая семантические аспекты.

Эффективное удаление шумов и ненужных элементов в тексте, также входящее в этап предобработки, способствует более точному выделению ключевых понятий и связей между ними в графе. Этот процесс помогает сфокусироваться на существенных элементах текста, предоставляя более чистый и информативный контекст для графового анализа.

Важным этапом адаптации текстового анализа для графовых методов является извлечение именованных сущностей и их представление в виде вершин графа. Это позволяет выделить ключевые элементы текста, такие как имена, места и даты, и представить их как важные узлы

графа. Дополнительное преобразование извлеченных сущностей через их категоризацию вносит дополнительные уровни информации в граф. Категоризация позволяет классифицировать сущности в соответствии с их семантическим значением, что, в свою очередь, углубляет анализ и предоставляет дополнительные возможности для классификации и интерпретации информации.

Word Embeddings, или векторные представления слов, предоставляют дополнительный уровень абстракции к анализу текста. Векторные представления слов, в сущности, представляют собой числовые векторы, где каждое слово кодируется в многомерном пространстве. Эта кодировка основывается на семантических связях и контекстуальных отношениях между словами. Важно подчеркнуть, что применение этих векторов в графовом анализе текста позволяет эффективно учесть семантическую близость и расстояние между терминами. Включение векторных представлений слов в граф обогащает его семантической информацией, предоставляя численные веса для рёбер. Это позволяет более точно выделить ключевые концепции, учитывая контекстуальные связи между словами.

Использование векторных представлений слов не ограничивается только улучшением точности весов рёбер в графе. Эти векторы могут быть интегрированы в различные алгоритмы, такие как обнаружение общностей, ранжирование, классификация и кластеризация узлов. [330] Таким образом, векторные представления слов предоставляют богатый источник информации для различных аспектов анализа текстовых данных в контексте графов.

Извлечение ключевых слов с использованием графовых методов представляет собой эффективный способ выделения наиболее значимых терминов в тексте и анализа их взаимосвязей. В данном случае граф структурируется таким образом, что каждый узел представляет собой отдельное ключевое слово, а связи между узлами отражают степень их семантической близости или взаимосвязи в тексте, т.е., каждый узел может быть взвешен в соответствии с его значимостью. Таким образом, методы, основанные на графах, не только выделяют ключевые слова, но и выявляют структуру и взаимосвязь между ними, что способствует более глубокому пониманию текста.

Этому посвящена статья “Double Graph Based Reasoning for Document-level Relation Extraction”.[434] Авторы описывают новые механизмы по извлечению отношений между сущностями на уровне документа. В отличие от извлечения отношений на уровне предложения, данная задача требует рассмотрения нескольких предложений внутри документа. В работе предложена модель Graph Aggregation-and-Inference Network (GAIN) с использованием двойного графа. GAIN сначала строит граф на уровне упоминаний для моделирования взаимодействий между различными упоминаниями в документе. Также создается граф на уровне сущностей, на основе которого предложен новый механизм осмыслиния для выявления отношений между сущностями.

Тематическое моделирование представляет собой метод, позволяющий выявить темы, которые присутствуют в коллекции текстов. Это особенно полезно при анализе больших объемов данных, таких как социальные медиа или новостные ленты, где разнообразие тем может быть велико. Каждая тема может быть представлена как узел графа, а связи между темами — как рёбра. Такой графовый подход отражает структуру тематических отношений в тексте. Узлы, представляющие темы, могут быть взвешены в соответствии с их важностью или распространённостью в коллекции текстов, а связи между ними могут отражать степень взаимосвязи или схожести. Преимущество тематического моделирования в контексте графов заключается в том, что оно не только представляет перечень тем, но и их взаимосвязи и структуру, что дает более глубокое понимание того, как различные темы взаимодействуют друг с

другом внутри текстовой коллекции.

Интеграция алгоритмов анализа тональности с графовыми методами открывает новые возможности для более глубокого понимания контекста текста. Графовый подход к анализу тональности не только выявляет эмоциональные компоненты в тексте, но также структурирует их в виде взаимосвязанных элементов. Например, в статье “A Generic Graph-Based Method for Flexible Aspect-Opinion Analysis of Complex Product Customer Feedback” [213] описывается универсальный графовый метод для гибкого анализа аспектов и мнений в сложных обзорах потребителей о продукции. В большинстве традиционных решений часто требуется использование размеченных наборов данных, они ориентированы в основном на классификацию тональности, не интегрируют области знаний, уступая по вариативности и гибкости в анализе аспектов и мнений. Предложенная модель поддерживает более широкий спектр аналитики в рамках единой структуры без повторной предобработки данных и нового моделирования. Конкретно данный метод позволяет проводить как обычный, так и сравнительный анализ аспектов и мнений, ранжирование удовлетворенности и влияния аспектов, извлечение тенденций в мнениях и преднастроенную (таргетированную) суммаризацию аспектов и мнений.

Адаптация классических алгоритмов анализа графов становится следующим важным этапом. Это включает в себя приспособление алгоритмов PageRank и community detection для работы с графиками, построенным на основе текстовых данных. Эти алгоритмы могут быть применены для выделения важных узлов, обнаружения сообществ и анализа структуры графа текста. Адаптация PageRank к таким графикам позволяет выявлять важные слова или текстовые элементы в контексте текста, отражая их влияние на структуру и смысловую связность текста. Например, в статье [256] авторы предлагают метод графовой суммаризации текста, основанный на модифицированном алгоритме TextRank, который учитывает важность предложений в документе. Метод создает граф, где узлы представляют собой предложения, а веса ребер определяются схожестью между предложениями. Используется модифицированная обратная частота предложения и косинусная схожесть для различной значимости слов в предложении. Граф разбивается на кластеры, предполагая, что предложения внутри кластера похожи друг на друга, что делает метод более эффективным для извлечения ключевой информации из текста.

В контексте вызовов современного информационного общества, где объемы данных постоянно увеличиваются, технологии также развиваются, предлагая все более эффективные методы для обработки и понимания текстовых данных. Объединение методов текстового анализа и графовых структур открывает перспективы для глубокого и комплексного понимания текста в разнообразных областях. В ходе проведенного исследования были рассмотрены различные аспекты взаимодействия графовых методов и текстового анализа в обработке больших объемов текстовых данных. Применение и эффективность этих методов на практике освещены в многочисленных исследовательских работах, где успешно применялись графовые методы для анализа текстов в областях от социолингвистики до оценки отзывов о продукте. Таким образом, интеграция графовых методов и текстового анализа предоставляет эффективные инструменты для более глубокого понимания контекста текстовых данных вне зависимости от доменной области. Графы позволяют визуализировать и анализировать сложные связи между элементами текста, тогда как текстовый анализ и методы компьютерной обработки текстов дают мощный инструментарий для более эффективного извлечения смысла и информации из текстовой информации. Таким образом, адаптация методов текстового анализа для графовых методов представляет

собой перспективное направление для более глубокого и комплексного анализа текстовых данных и максимизации их информационного потенциала, что может привести к более эффективным методам анализа и принятия решений в различных областях как в науке, так и в бизнесе.

## 7 Сравнительный анализ актуальных подходов к анализу неструктурированной текстовой информации (стохастический блокмоделинг, LDA и BERT модели) на примере анализе дискурсов в социальных медиа

### 7.1 Введение

Повсеместное использование социальных сетей в качестве средства коммуникации подчеркивает необходимость внимания к особенностям методологии и инструментария, необходимых для анализа данных, производимых пользователями социальных сетей. В частности, актуализируется вопрос обработки коротких текстов, недостаток семантического контекста в которых может стать серьезным ограничением для работы многих алгоритмов машинного обучения, применяемых к текстовым данным.

Одной из особенностей текстовых данных, производимых в социальных сетях, является отсутствие «разметки» - не существует истинных меток тематической принадлежности, тональной нагрузки и прочих характеристик содержания текстового источника. В таких случаях повышается релевантность методов машинного обучения «без учителя» (unsupervised), способных идентифицировать сходные группы внутри данных. Тематическое моделирование – направление в рамках дисциплины обработки естественных языков (NLP), направленное на агрегацию текстов в тематические кластеры. Практические приложения методов тематического моделирования кросс-дисциплинарны и включают рекомендательные системы [godin2013?], научометрию [asmussen2019?], дискурсивные исследования [lyu2021?, hoseini2023?], автоматической идентификации событий в новостях и социальных сетях [hu2012?, qian2015?, gong2017?, lyu2021?] и проч. Тем не менее, предпосылки о данных, на которых основаны классические алгоритмы тематического моделирования, не всегда выполняются для текстов сообщений в социальных сетях. Короткая длина текстов, мотивируемая платформенными лимитами на объём публикаций и комментариев, не предоставляет достаточного семантического контекста для анализа слов классическими методами [hong2010?, albalawi2020?, quiang2022?].

Следовательно, необходимо формализованное сравнение качества разных алгоритмов тематического моделирования для коротких текстов с целью обнаружения наиболее эффективного с вычислительной и интерпретационной точки зрения подхода. В отличие от большинства обзорных исследований по теме, мы фокусируемся не на анализе вариаций латентного распределения Дирихле (LDA), а предлагаем сравнить его работу с методами, основанными на стохастическом блокмоделировании и векторных представлениях слов в модели типа «трансформер», соответственно.

Разнообразие сравниваемых методов позволяет более конкретно изучить связь между устройством подхода к тематическому моделированию и особенностями результирующих тематических кластеров, а также отметить значимые направления потенциального развития и оптимизации наличествующих инструментов для задач анализа коротких текстов.

## 7.2 Цели исследования

Провести формальную оценку качества алгоритмов тематического моделирования на данных коротких текстов на примере комментариев в социальной сети TikTok. Оценить плюсы и недочеты разных инструментов тематического моделирования в приложении к коротким текстам, производимых в социальных сетях.

## 7.3 Практическая значимость исследования

Настоящее исследование призвано проанализировать эффективность алгоритмов, относящихся к разным подходам в тематическом моделировании, на нетипичном наборе данных: в отличие от большинства датасетов, зачастую используемых для оценки методов тематического моделирования, комментарии в социальных сетях содержат большое количество опечаток и ошибок, сокращений, неконвенциональной лексики (например, сленга), при этом имея крайне короткую длину. В то же время, понимание перспектив и ограничений использования методов тематического моделирования именно для данных социальных сетей может стать важным инструментом для исследований медиа-коммуникации разной дисциплинарной и тематической направленности.

## 7.4 Обзор современных методов тематического моделирования

Методы тематического моделирования получили широкое распространение в конце 1990-х годов с появлением подхода PLSA [hofmann2013?] и его последующей адаптации, ставшей конвенцией в тематическом моделировании – LDA [48]. Методы тематического моделирования обычно основаны на некоторых предположениях о взаимосвязи между темами и документами. В частности, один из наиболее распространенных алгоритмов тематического моделирования LDA предполагает, что каждый документ представляет собой смесь тем, а каждая тема – распределение Дирихле по словам. Распределение Дирихле является многомерной формой бета-распределения, которое используется для генерации распределений для каждого слова в наборе данных. В целом генеративная тема модели смешения Дирихле-Мультиномиального распределения имеет такую структуру:

Для каждой темы  $k \in 1, \dots, K$   $\phi_k \sim Dir(\beta)$   $\theta \sim Dir(\alpha)$

Для каждого документа  $m \in 1, \dots, M$   $z_m \sim Mult(1, 0)$  и для каждого слова  $n \in 1, \dots, N_m$   $x_{mn} \sim Mult(1, \phi(z_m))$ , где

К – количество тем М – количество документов в датасете N – количество слов в документе

Хотя модели такого типа продемонстрировали высокую эффективность и качество для анализа художественной и научной литературы, что сделало их стандартом в отрасли, их применение для коротких текстов, доминирующих в социальных сетях, ограничено недостаточностью семантического контекста в коротких текстах, а также спорностью предпосылки наличия распределения тем внутри документа поскольку многие из них содержат только одну тему [ahuja2015?, quiang2022?].

Решение этой проблемы предлагают [dieng2020?], которые считают, что классическая LDA может достигать более высоких результатов в задачах тематического моделирования за счет комбинации с эмбеддингами слов (авторы используют word2vec skip-gram, но технических ограничений на то, как могут выглядеть векторные представления слов, нет). Предложенная ими модель ETM моделирует каждое слово категориальным распределением, естественным параметром которого является произведение между

эмбеддингом слова и эмбеддингом назначеннной ему темы, что позволяет сохранять и использовать больше семантической информации о словах в документах. Возможность ссылаться на общий контекст слова очень важна для эффективной обработки коротких документов, поскольку семантический контекст, который может быть выведен из самого набора данных, может быть искажен. Поэтому для получения более качественных кластеров слов наряду с традиционным LDA можно использовать предварительно обученные эмбеддинги.

Среди прочего, короткие текстовые данные предлагается обогащать метаданными, тэгами авторов [rosen-zvi2012?] и хэштегами [wang2016?] использовать длинные тексты для обучения модели тематического моделирования для инференса на коротких текстах [phan2008?] и проч. Эти адаптации, однако, не нивелируют проблему отсутствия критерия выбора количества тем (хотя в этом направлении также ведется работа, см. [koltcov2021?]) и обоснования предпосылки о соответствии распределения тем и слов внутри темы распределению Дирихле [gerlach2018?].

Вместо семантического обогащения модели [gerlach2018?] предлагают структурное изменение, подходя к задаче тематического моделирования как к задаче выделения сообществ в сетях. Они предлагают метод иерархического стохастического блокмоделирования (hSBM) как более эффективный и адаптируемый алгоритм для выделения тем. Стохастические блочные модели – это статистический фреймворк, используемый для моделирования сложных сетей. В этой модели узлы сети разбиваются на различные группы или блоки, при этом вероятность появления ребер между узлами зависит от принадлежности узлов к той или иной группе. В модели предполагается, что узлы одной группы имеют схожий характер связности, а узлы разных групп - разный характер связности. Вероятность возникновения ребра между двумя вершинами определяется принадлежностью их к кластерам, причем вероятность возникновения ребра между вершинами одного кластера выше, чем между вершинами из разных кластеров. В основе hSBM лежит идея замены неинформативных приоров, используемых в LDA, на непараметрический подход: глубокую байесовскую иерархию приоров и гиперприоров, которые не делают предположений о свойствах данных более высокого порядка. Таким образом, авторы делают вывод о том, что hSBM могут стать решением вышеупомянутых недостатков традиционного LDA. Более того, они подтверждают свое утверждение, сравнивая производительность своей модели с LDA на большом наборе данных статей Википедии, а также искусственно созданных текстовых наборов данных, не соответствующих основным предположениям LDA, где hSBM значительно превосходит LDA (Там же).

Похожая логика применения иерархичности в выделении групп для сохранения и отображения гетерогенности в данных применяется и в наработках среди моделей тематического моделирования, ориентированных на расширение семантического контекста анализируемых документов. Например, методы, использующие представления слов, создаваемые трансформерными моделями, прибегают к иерархической кластеризации эмбеддингов как способу выделения тем. [grootendorst2022?] описывает процедуру выделения тем при помощи трансформеров следующим образом:

1. Для создания эмбеддингов документов используется предварительно обученная языковая модель на основе трансформера;
2. Для учета недостатков высокой размерности вкраплений представления документов уменьшаются с помощью Uniform Manifold Approximation and Projection (UMAP) - инструмента для уменьшения размерности, превосходящего t-SNE и PCA за счет более высокой скорости сохранения локальных и глобальных особенностей данных [mcinnes2020?];

3. Редуцированные вкрапления кластеризуются с помощью HDBSCAN - иерархического расширения классического DBSCAN, позволяющего выявлять кластеры различной плотности. При мягкой кластеризации шум помечается как выбросы, которым не присваивается кластерная метка, чтобы избежать включения несвязанных данных в какую-либо из тем;

4. Для определения важности слов по кластерам, а не по документам, используется классовый TF-IDF, который позволяет получить распределения слов по темам для всех тематических кластеров. Векторы с-TF-IDF нормируются с помощью L1-нормы для сглаживания представлений тем.

Описанные особенности модели делают BERTopic высокопотенциальным подходом к тематическому моделированию, поскольку представления тем являются динамичными и гибкими, а также восприимчивыми к временным изменениям текстовых данных.

Так, развитие области тематического моделирования отмечается расширением спектра типов данных, с которыми алгоритмы выделения тематических групп способны эффективно работать. Хотя наиболее популярным аналитическим направлением стало обогащение контекста коротких текстов за счет предварительно обученных вкраплений и других типов данных, подключаемых к модели, были сделаны шаги и в сторону изменения концептуальных основ подхода к задаче тематического моделирования, одним из которых является использование тематических кластеров, произведенных методами сетевого анализа.

## 7.5 Методология и дизайн исследования

Целью настоящего исследования было сравнение качества алгоритмов тематического моделирования для анализа коротких текстов, представленных на примере комментариев и хэштегов в социальной сети TikTok. Мы прибегаем к сравнению алгоритмов, относящимся к разным группам методов. Опираясь на категоризацию, предложенную [abdelrazek2022?], мы используем модели из двух разных из выделенных ими групп: вероятностные (LDA и SBMTM), нейросетевые (ETM и BERTopic).

LDA обогащается предобученными эмбеддингами GloVe (Global Vectors) для русского языка Navec<sup>16</sup>. Выбор обоснован малым объёмом памяти, необходимой для их использования и дообучения при большом объёме словаря: используемые эмбеддинги hudlit\_12B\_500K\_300d\_100q покрывают 98% слов в художественных текстах, занимая 95.8 Мб RAM. GloVe основан на глобальном статистическом подходе: целью обучения является минимизация разницы между точечным произведением векторов слов и логарифмом вероятностей соприсутствия в корпусе на основе матрицы соприсутствия. Использование глобальных статистик совстречаемости слов позволяет определить сравнительную значимость слов в тексте [pennington2014?].

Для триангуляции оценки качества, описанные алгоритмы сравниваются при помощи следующих метрик:

- Время, затрачиваемое на вычисления;
- Когерентность (связность) выделяемых тем

Оценка связности производится при помощи метрик NPMI и UMass на топ-10 словах каждой темы: эти метрики рассчитывают близость слов в теме на основе паттернов их совместного появления в корпусе. NPMI хорошо коррелирует с человеческой оценкой [aletras2013?], однако UMass более устойчива к вариациям в референтном корпусе и шуму, а также, в отличие от NPMI, не агностична к позиции слова

<sup>16</sup><https://natasha.github.io/navec/?ysclid=l9kgcwq6eu573989592>

при анализе: значение метрики будет меняться в зависимости от того, какое слово является центральным, а какое – контекстным при расчёте совстречаемости.

- Разнообразие тем

Мы используем метрику, обратную Rank-Biased Overlap (RBO) – рекурсивно оценивающую долю пересечений между ранжированными по значимости списками слов в теме на разных уровнях «глубины» погружения в список. Низкая доля подобных пересечений между выделяемыми темами указывает на их высокое разнообразие.

## 7.6 Эмпирическая база исследования

Для задачи оценки качества на неразмеченном датасете мы используем массив русскоязычных комментариев к урбанистическим видео в TikTok. Выбор TikTok в качестве платформы для анализа обусловлен устойчивой популярностью сети до блокировки в России, а также адаптацией формата в прочих социальных сетях, до сих пор доступных на территории РФ (YouTube, ВКонтакте). Фокус на урбанистической тематике основан на потребности в общей тематической когерентности анализируемых видео, которую можно использовать в качестве базового бенчмарка тематики. Сама по себе тема урбанистики удобна своей популярностью, предоставляющей достаточно объёмную базу для исследования, а также способностью вовлекать аудиторию в дискуссию за счёт близости темы зрителям, что обеспечивает наличие содержательных обсуждений в комментариях.

Отбор видео осуществлялся по хэштегам, список которых расширялся при помощи рекомендательного алгоритма платформы. Финальный набор содержит 17 позиций. Из массива видео, относящихся к этим хэштегам, доступными оказались 2625 видео и содержащими комментарии - 1271.

Хэштег	Кол-во видео	Кол-во просмотров
человейники	368	8000000
метроспб	292	84600000
московскоеметро	249	89900000
городская среда	243	13800000
транспорт	214	448600000
урбанизм	207	36600000
реконструкция	198	21400000
благоустройство	189	116300000
монорельс	183	4100000
общественный транспорт	179	77800000
урбанистика	170	74900000
часник	93	111600000
комфортная среда	87	456800
урбанизация	46	614800
транспортная реформа	33	183000
безбарьерная среда	21	1200000
городское планирование	5	381000

Выгрузка комментариев производилась с помощью Selenium WebDriver для Chrome и языков программирования JavaScript и Python.

Тексты комментариев были предобработаны при помощи приведения всех слов-токенов к нормальной форме, удалением стоп-слов и фильтрацией токенов, встречающихся менее 10 раз или в более, чем 80% документов. В анализ включались единичные токены, а также биграммы. Для предобработки использовались библиотеки NLTK, pymorphy2 и Gensim для Python.

## 7.7 Результаты

Мы обнаруживаем, что, с точки зрения связности результирующих тематических кластеров, ни одна из используемых моделей не демонстрирует результатов высокого качества, для большинства из них качество можно назвать средним (Табл. №2). Если по метрике NPMI результаты работы алгоритмов сходны, то по метрике связности UMass расхождения велики: SBMTM демонстрирует более низкое качество, что указывает на более низкое качество ранжирования слов внутри темы и недостаток точности при агрегации слов в кластеры. Чем выше уровень иерархии, тем ниже значение UMass – при объединении тем их связность падает. Опираясь на диапазоны значений NPMI, мы также можем заметить, что качество тем сильно различается внутри каждой из моделей: они содержат как темы с низкой, так и с более высокой когерентностью.

	SBMTM	SBMTM	ETM	BERTopic
	Level 1	Level 2	Navec	
Время на вычисления	244	244	128	46
Количество тем	11	2	60	19
NPMI (диапазон значений)	[-0.19;0.21]	[-0.05;0.03]	[-0.117; 0.311]	[-0.05;0.339]
Средний NPMI по темам	0.05	-0.01	0.042	0.12
Средний UMass по темам	-7.66	-8.7	-3.96	-0.23

Более того, следует обратить внимание на то, что разброс значений когерентности тем у SBMTM выше, чем у BERTopic: если последний инструмент склонен создавать темы примерно одного и того же среднего уровня качества, то SBMTM производит большое количество как очень низко, так и очень высоко когерентных тем.

Распределение когерентности тем BERTopic и SBMTM

Figure 18: Распределение когерентности тем BERTopic и SBMTM

С точки зрения разнообразия тем ETM уступает BERTopic (для SBMTM метод, производящий пересекающиеся сообщества, не имплементирован): в составе тем, производимых ETM, присутствует много повторяющихся высокочастотных слов (например, «человек», «общество», «город»), в то время как BERTopic производит более уникальные по составу темы. Результаты работы ETM далее не

интерпретируются ввиду низкой уникальности тем и плохой различительной способности слов, входящих в итоговые темы.

BERTopic значительно превосходит свои альтернативы по всем выбранным критериям качества модели - он самый быстрый и выдаёт наиболее интерпретируемые и понятные темы по сравнению с конкурентами. Такая существенная разница в качестве может быть частично объяснена контекстуальными эмбеддингами, используемыми в BERT, то есть эмбеддинги слов различаются от предложения к предложению в зависимости от их контекста, а также структурного положения в предложении. Такой подход может быть наиболее полезен для тематического моделирования коротких текстов, поскольку статистические эмбеддинги могут не распознать устойчивые паттерны соприсутствия или семантической связи из-за малой длины документа и, соответственно, небольшого размера окна для обучения.

В целом, в соответствии с утверждениями автора оригинальной статьи, BERTopic, как с редукцией тем, так и без нее, создает хорошо интерпретируемые темы. Однако эти темы чрезвычайно специфичны и контекстуальны. Другими словами, BERTopic выбирает узкие темы, а не более широкие, и чувствителен к локальным паттернам дискурсивной активности.

Топ-5 слов для выделенных BERTopic тем

Figure 19: Топ-5 слов для выделенных BERTopic тем

Для SBMTM на первом уровне агломерации модель выделила 11 тем, которые представлены в таблице №3 вместе с 10 наиболее частотными примерами слов каждой из них. Помимо тематических тем, модель также группировала слова по их языку, если он не русский, и по структурной позиции в предложении. Так, темы 10 и 11 представляют английские и украинские слова, а тема 3 состоит из слов речевых актов, таких как приветствия, поздравления, благодарности и однословных экспрессивных высказываний: интересно ('интересно'), круто ('топ', 'здраво'). Остальные темы сгруппированы в зависимости от предмета обсуждения.

Темы 1 и 5 трудно интерпретировать однозначно, так как они состоят из разных слов, обозначающих городскую жизнь, где есть упоминания и о транспорте, и о работе, и о деньгах, и т.д. Тема 2 может быть интерпретирована как единичное высказывание, так как все наиболее часто встречающиеся слова в этой теме относятся к одной и той же идеи - жители Москвы сообщают автору публикации о том, что, несмотря на проживание в Москве, они не знают о чем-то, что упоминается в видеоролике. Тема 4 посвящена общественному транспорту в целом и, в частности, в Санкт-Петербурге.

Темы 6-9, однако, более тематически целостны и, следовательно, могут быть интерпретированы однозначно. Тема 6 касается конкретного вопроса пользования общественным транспортом - уступать ли свое место другому пассажиру (в частности, женщинам, детям и пожилым людям). Примечательно, что слово, обозначающее пожилую женщину, носит сатирический характер: вместо "бабушка" используется "бабка", что считается неуважительным описанием, и подчеркивает, что вопрос вызывает споры и негативные настроения. Тема 7 включает в себя общепотребительную лексику, связанную с использованием общественного транспорта. Наиболее употребительные слова этой темы отражают правила и ожидания людей, пользующихся общественным транспортом, а также слова, указывающие на возможные проблемы и исключения. Например, слово "сторона" (side) часто используется в наборе данных наряду со словом "лифт" для обсуждения вопроса о том, предпочтительно ли пользоваться

обеими сторонами лифта или оставлять левую сторону открытой для желающих подняться по лестнице - весьма распространенная практика в Москве, однако не соответствующая правилам пользования метрополитеном. Другой пример - слово "коляска" (может означать детскую или инвалидную коляску), которое также может относиться к проблемам, с которыми сталкиваются маломобильные люди или родители при пользовании общественным транспортом. Тема 8 посвящена жилым домам, включая упоминания конкретных типов зданий, таких как "хрущёвка" (хрущёвка - тип советских жилых домов, массово выпускавшихся по плану обеспечения большинства населения доступным жильём), а также иронические названия высотных зданий ("человейник" - игра слов, сочетающая "человек" и "муравейник", чтобы подчеркнуть, насколько тесны некоторые из новых жилых комплексов). Тема 9 включает политизированную лексику, сравнивающую Россию с ее предполагаемыми политическими противниками.

Тема №	Топ-10 слов
1	город, год, делать, время, деньги, работать, район, дорога, жизнь, хотеть
2	жить, Москва, знать, видеть, сделать, хороший, думать, видео, смотреть, первый
3	спасибо, привет, молодец, интересно, круто, жиза, топ, классно, поздравлять, здорово
4	метро, автобус, трамвай, ездить, станция, новый, Питер, поезд, ветка, ходить
5	человек, говорить, сказать, стоить, машина, работа, идти, проблема, понимать, дело
6	место, ехать, ребенок, уступать, сидеть, женщина, девушка, мужчина, бабка
7	стоять, дверь, сторона, остановка, выходить, правило, эскалатор, быстрый, переход, коляска
8	дом, строить, квартира, построить, купить, человейник, земля, покупать, хрущёвка, этаж
9	Россия, страна, Путин, бог, Европа, Украина, русский, слава, Америка, завидовать
10	You, the, that, and, not, just, now, what, peace, she
11	расти, укр, вонь, петров, Крым, але, така, буде, кто, немой

Выше по иерархии тем группы второго уровня состоят только из двух кластеров, объединяющих описанные выше темы. На втором уровне иерархии тем большинство описанных выше тем объединяются в группу слов, относящихся к городской жизни, включающую как темы жилых домов, так и общественного транспорта. Примечательно, что среди наиболее частотных слов в этой теме встречается топоним - название столицы России, Москвы. Вторая тема, обособленная от урбанистических дискуссий, содержит политизированную лексику.

Тема №	Топ-10 слов
1	жить, человек, город, метро, дом, год, Москва, место, автобус
2	Россия, страна, Путин, бог, Европа, Украина, русский, слава, завидовать

## 7.8 Обсуждение и выводы

Результаты анализа качества тематического моделирования с использованием метрик coherence и разнообразия тем показывают, что BERTopic превосходит альтернативные методы в создании когерентных тем. Преимущество этого алгоритма достигается благодаря использованию c-TF-IDF, которое является уникальным алгоритмом взвешивания слов по их значимости внутри темы и отсутствует у других методов. Однако, было обнаружено, что хотя темы, выделенные BERTopic, имеют в среднем более высокую когерентность, SBMTM на нулевом уровне иерархии создает больше тем с высокой когерентностью, что уравновешивается темами с низкими показателями качества при усреднении. Применение совместного моделирования сообществ для документов и слов в бимодальной сети позволяет SBMTM более эффективно кластеризовать документы, однако его темы более подвержены генерализации. Иерархическая структура SBMTM успешно вносит гетерогенность в производимые темы, в то время как BERTopic сконцентрирован на локальных семантических паттернах.

Проведенные эксперименты подтверждают перспективность использования альтернативных методов LDA для тематического моделирования на коротких текстах, особенно сетевого подхода к представлению текстовых данных. Мы находим, что иерархическое блокмоделирование превосходит методы, основанные на словарных эмбеддингах, по интерпретируемости и генерализации интерпретаций, что было продемонстрировано ранее в задаче классификации документов на англоязычных датасетах [qiang2022?]. Тем не менее, следует отметить, что SBMTM уступает альтернативам по производительности и качеству ранжирования слов в темах по значимости. Это подтверждается высоким разбросом значений и заметными различиями в распределении coherence в зависимости от числа анализируемых слов. Таким образом, сетевой анализ представляет потенциал для тематического моделирования на коротких текстах, однако методы, основанные на более сложных внутренних представлениях текста, демонстрируют более высокое качество путем лучшей репрезентации близости между текстами и результирующими темами.

## 7.9 Заключение

Мы сравнили три инструмента для тематического моделирования: иерархическое стохастическое блокмоделирование (SBMTM), Latent Dirichlet Allocation с эмбеддингами слов Navec и подход на основе трансформер-модели (BERTopic) - и пришли к выводу, что последний значительно превосходит свои альтернативы на нашем корпусе. Тем не менее, несмотря на вычислительное преимущество, BERTopic может быть нежеланным инструментом для исследователя ввиду того, что производимые им темы крайне локальны и не позволяют выделять более генерализуемые направления в дискурсе. SBMTM не страдает от этой проблемы и производит более генерализуемые темы, однако производит больше тем с низкой когерентностью.

## **7.10 4.2.1 Библиометрический сетевой анализ коллaborаций российских социологов на материалах Web of Science**

Включенность в международное исследовательское сообщество является важной предпосылкой становления и развития исследовательских институтов и научных школ. Конкурентоспособность научных коллективов, работающих в рамках любой научной дисциплины, во многом зависит от сотрудничества как внутри страны, так и на международном уровне.

Библиометрические исследования на протяжении всей истории их существования фиксируют тенденцию к увеличению научной коллаборации [352], в том числе коллаборации международной. Такой анализ не только позволяет узнать, каким будет облик науки будущего, но и оценить рост влияния совместной работы ученых на перспективы того или иного научного сообщества. Все это делает изучение коллабораций ученых в контексте влияния их публикаций крайне актуальным. Наиболее распространенный и эффективный метод анализа научных коллабораций – сетевой анализ сетей соавторства [251], данные для которых могут быть получены из наукометрических баз данных [308] (Scopus, WoS, eLibrary и др.).

И.Н. Трофимова проанализировала на основе базы данных Web of Science публикации российских ученых с 2018 по 2022 годы, фокусируясь на международной коллаборации российского научного сообщества. В частности, она отмечает рост числа публикаций, написанных российскими учеными в соавторстве с иностранными коллегами, происходящий на фоне снижения влияния российских публикаций в мировых масштабах [482]. Также данное исследование подтверждает положительную связь между международным соавторством и цитируемостью публикаций (треть цитируемых публикаций российских ученых написаны в соавторстве с иностранцами, эти публикации чаще выходили в журналах Q1). География международного соавторства российских ученых в большей степени определяется исторической развитостью научных центров и объемами финансирования, чем территориальным расположением государств и культурной близостью (наибольшее число иностранных соавторов российских ученых из США, стран Европы, Китая и Японии, а научное сотрудничество со странами СНГ менее продуктивно).

Х.Ф. Моэд, В.А. Марсукова и М.А. Акоев провели сравнительное исследование трендов публикационной активности российских ученых на основе баз библиометрических данных Web of Science и Scopus [270]. Анализ показал сильную разницу в оценках роста числа российских публикаций и их влияния. Авторы исследования пришли к выводу, что на положительную динамику российского научного вклада “наложились” изменения числа русскоязычных изданий, включенных в базы Web of Science и Scopus, что вызывает трудности в ее оценке и говорит о необходимости учитывать особенности каждой из баз данных для построения валидных выводов при работе с ними.

В практике оценки продуктивности научного сообщества растет важность как самих методов наукометрического и библиометрического анализа, так и баз данных (БД), которые являются основными поставщиками библиографических метаданных о публикационной активности исследователей. Как правило, библиографические базы данных Web of Science (WoS) и Scopus определяются в качестве наиболее полных источников данных для различных аналитических целей [442]. Несмотря на то, что две основные специализированные БД – WoS и Scopus – по-прежнему считаются наиболее надежными источниками библиографических данных, именно WoS все же рассматривается как «золотой стандарт»

библиометрического использования [313].

По данным российской наукометрической базы eLibrary, средняя цитируемость статей, опубликованных в WoS, включенных в ядро Российского индекса научного цитирования (РИНЦ), отличается от аналогичных статей из Scopus – в WoS она в 1,25 раз выше, чем у статей в Scopus, в 9,3 раза выше, чем в ESCI (Emerging Sources Citation Index), в 6,7 раз выше, чем в RSCI (Russian Science Citation Index). При распределении на квантили средняя цитируемость статей в WoS Q1 в 1,36 раз выше Scopus Q1, в 28,2 и 20,3 раза выше, чем в ESCI и RSCI соответственно. Средняя цитируемость российских статей в WoS Q4 в 2,7 раз выше, чем в Scopus Q4, в 4,4 и 3,1 раза выше, чем в ESCI и RSCI соответственно [111]. Преимущества базы данных WoS как с точки зрения публикующихся авторов, так и с точки зрения качества метаданных делают ее наиболее подходящим источником для получения валидных результатов библиометрического анализа.

Настоящее исследование посвящено библиометрическому анализу международного измерения публикаций российского социологического сообщества на основе материалов базы Web of Science за 1992-2022 гг. Основной акцент в работе сделан на сетях соавторства в интересах выявления уникальных паттернов коллабораций российских социологов через публикации, индексируемые в наукометрической базе WoS. Подмножество по социологии было выделено на основе категории SC (SC = Sociology, категория исследования/research area в WoS) и состоит из 7915 публикаций со спецификацией поля данных «CU=(Russia)» в режиме full record.

Всего было проанализировано 7915 публикаций всех типов (статьи, монографии, конференционные публикации и др.) из 172 изданий, опубликованных за период с 1992 по 2022 (май). Для целей данного исследования авторы не исключали никакие публикации из набора данных для того, чтобы все публикации, проиндексированные в WoS, попали в анализ. Более 40% публикаций не имели соавторов. Такое значительное количество работ без соавторов предполагает, что в российском социологическом сообществе довольно значительное число авторов работает индивидуально, не вступая ни в какие коллаборации. На каждую публикацию в среднем приходилось 1,57 соавтора, 1,419 цитирований, а возраст всех публикаций в среднем составил 12,7 лет. Доля международного соавторства составила 4,611%.

Зачастую для библиометрических исследований сетей характерно применение комбинаций программных продуктов, которые можно использовать алгоритмически комплементарно друг другу. Выбор комбинаций программ для анализа часто также зависит от исследовательского вопроса. В данной работе построение библиометрических сетей и проведение библиометрического анализа осуществлялись при помощи нескольких программных продуктов – VOSviewer, Pajek и R (библиотека bibliometrix/biblioshiny).

В анализе цитирования представлены как зарубежные, так и российские исследователи. Это предполагает, что в российском социологическом сообществе сформировались отечественные научные школы, заметные международному исследовательскому сообществу. Однако список топ-журналов, где публикуются российские социологи, довольно ограничен. На Рис. 20 представлена сеть цитирований с источниками (изданиями), где публикуют работы эти авторы. На первом месте с большим отрывом стоит журнал «Социологические исследования».

В Таблице ?? приведены данные по топ-10 изданиям по показателям общего количества публикаций, цитированию и общей силы связей (total link strength). Обращает на себя внимание, что

Сеть цитирований по источникам (изданиям) публикаций

Figure 20: Сеть цитирований по источникам (изданиям) публикаций

в топ-10 источников по количеству опубликованных документов вошли международные конференции, которые готовили публикационные материалы как журналы (по факту те же самые сборники конференционных материалов) – Social and Cultural Transformations in the Context of Modern Globalism (European Proceedings of Social and Behavioural Sciences) или International Multidisciplinary Scientific Conference on Social Sciences and Arts SGEM 2016 (SGEM Conference Proceedings). Можно сделать вывод, что гонка 2010-х за продуктивностью международных публикаций, с которой пришлось столкнуться российскому академическому сообществу, нашла свое отражение в списке источников публикаций, где среди топ-источников оказались представленными сборники материалов конференций, которые не имеют эквивалентного репутационного веса по сравнению с академическими журналами. В этом контексте будет полезным сравнение не только по количеству опубликованных документов, но также по количеству цитирований и метрике «Общая сила связей» (total link strength). В VOSviewer для элемента сети учитывается количество связей элемента с другими элементами (links) и общую силу связей элемента с другими элементами (total link strength). Например, в случае сетей соавторства, авторы, имеющие одинаковое число соавторов, будут иметь один и тот же показатель связей (буквально, сколько у них было коллабораций). Если же один из них будет чаще публиковаться совместно с кем-либо, но число его соавторов будет неизменным, то показатель общей силы связи у него будет выше, чем у другого исследователя. Таким образом, показатель общей силы связи учитывает не только наличие совместных публикаций, но и интенсивность соавторства, что позволяет получить более точные выводы относительно статуса ученых в сети.

Table 22: Топ-10 источников публикаций

НПП	Журнал	Публикации, шт.	Журнал	Цитирования, шт.	Журнал	Общая сила связей
1	Социологические исследования	4922	Социологические исследования	5525	Социологические исследования	260
2	Вестник Российского университета дружбы народов. Серия: Социология	679	Social Indicators Research	690	Социологическое обозрение / Экономическая социология	129

НПП	Публикации,		Цитирования,		Общая сила связей	
	Журнал	шт.	Журнал	шт.	Журнал	
3	Экономическая социология	540	International Journal of Intercultural Relations	557	Вестник Российской Федерации по дружбе народов.	47
4	Социологическое обозрение	517	Социологическое обозрение	396	Социология и технологии	38
5	Социальные и культурные трансформации в контексте современного глобализма (конференция, European Proceedings of Social and Behavioural Sciences)	322	Экономическая социология	394	Current Sociology	19
6	Социология науки и технологий	195	Вестник Российского университета дружбы народов.	270	Comparative Sociology	16
7	Changing Societies & Personalities	85	Population and Development Review	181	International Journal of Sociology and Social Policy	12

НПП	Журнал	Публикации,		Цитирования,		Общая сила связей
		шт.	Журнал	шт.	Журнал	
8	International Multidisciplinary Scientific Conference on Social Sciences and Arts SGEM 2016 (Psychology and Psychiatry	78	Ethics and Racial Studies	115	American Sociologist	11
9	International Journal of Sociology and Social Policy	41	Annals of Tourism Research	106	Critical Sociology	9
10	Comparative Sociology	29	European Societies	103	Filosofija. Sociologija	7

На Рисунке 21 представлен топ списка самых цитируемых публикаций исследователей (локальное цитирование и глобальное цитирование). Глобальное цитирование означает общее количество цитирований, которое статья, включенная в коллекцию, получила из документов, проиндексированных в библиографической базе данных в целом. Среди глобальных цитирований в основном представлены статьи, опубликованные в международных журналах. Локальные цитирования получены по публикации «внутри коллекции» (массива данных). Среди локально цитируемых публикаций в основном представлены статьи из журнала «Социологические исследования».

Топ цитируемых публикаций (вверху – локальные цитирования, внизу – глобальные цитирования)

Figure 21: Топ цитируемых публикаций (вверху – локальные цитирования, внизу – глобальные цитирования)

Библиометрические сети соавторства являются одними из основных видов сетей в библиографическом анализе и выражают определенный тип взаимосвязей между элементами изучаемого нами пространства. Сеть соавторства, как гласит название, отражает связи совместного участия агентов (исследователей, организаций, стран) в производстве академических публикаций. В этой сети узлами выступают авторы, а связь между ними отражает частоту, с которой они совместно публиковали статьи. Благодаря рассмотрению сетей соавторства мы можем оценить структуру научной коллaborации, выявить ключевых и периферийных акторов этого процесса. Сеть соприсутствия ключевых слов позволяет нам картировать тематический ландшафт академического поля, выявить приоритетные и популярные темы

исследований, а также то, на что исследователи обращают меньше внимания, либо не обращали его вовсе. Технически это осуществляется благодаря подсчету частоты, с которой термины одновременно встречаются в обозначенном поле библиографических данных.

В случае обоих сетей связь между элементами оценивается благодаря совместному подсчету авторства либо ключевых слов. Этот подсчет может быть полным либо фракционным. При полном подсчете мы считаем, что каждая связь между узлами сети имеет вес, равный числу документов, которое они опубликовали вместе (сеть соавторства) либо где они встречались вместе (сеть соприсутствия ключевых слов). Например, если 5 авторов выпустили одну публикацию, вес каждой их связи друг с другом равен 1; либо если в одной публикации встречаются 5 ключевых слов, вес связей между ними также будет 1. При фракционном подсчете вес связей будет определяться обратно числу узлов. Теперь каждый из соавторов нашей публикации будет связан друг с другом с весом  $\frac{1}{n}$ , как и каждый термин будет связан с другим весом  $\frac{1}{n}$ . Разница кажется небольшой, однако использование полного подсчета, который делается во многих исследованиях по умолчанию, приводит к значительному (практически квадратичному) увеличению числа связей с ростом числа авторов, что существенно искажает реальную картину научных коллабораций [305]. В связи с этим, в нашем анализе мы используем фракционный подсчет для построения сетей соавторства.

Наш массив данных состоит из 6765 авторов и 1664 организаций. За весь изучаемый период только 28% ученых опубликовали 2 и более работ (3 и более – 14%, 4 и более – 8%). Далее представлен топ-10 авторов по числу публикаций, числу цитирований и метрике общей силы связей (Таблица ??). Активно публикавшиеся авторы также, в основном, имеют сильные связи с другими авторами. Среди наиболее цитируемых авторов, однако, немного исследователей из России. Это обстоятельство объясняется международной кооперацией – многие из представленных ниже активно цитирующихся международных авторов работали с российскими коллегами и публиковались в российских журналах, индексируемых Web of Science.

Table 23: Топ-10 представленных социологов

НПП	Автор	Публикации,		Цитирования,		Общая сила связей
		шт.	Автор	шт.	Автор	
1	Троцук И.В.	69	Инглехарт Рональд	575	Троцук И.В.	33
2	Тощенко Ж.Т.	53	Вельцель Кристиан	532	Зборовский Г.Е.	21
3	Кравченко С.А.	44	Делхи Ян	407	Голенкова З.Т.	21
4	Радаев В.В.	41	Ньютон Кеннет	397	Пузанова Ж.В.	21
5	Зборовский Г.Е.	36	Шмит Петер	393	Нарбут Н.П.	20
6	Пузанова Ж.В.	34	Давидов Эльдад	318	Тощенко Ж.Т.	17

НПП	Публикации,		Цитирования,		Общая сила связей	
	Автор	шт.	Автор	шт.	Автор	
7	Барсукова С.Ю.	34	Берри Джон	214	Игитханян Е.Д.	14
8	Лапин Н.И.	33	Барсукова С.Ю.	154	Коротаев А.В.	13
9	Горшков М.К.	30	ван де Вийер Фонс	125	Ларина Т.И.	13
10	Голенкова З.Т.	28	Кравченко С.А.	124	Иванов В.Н.	13

Паттерн представленности организаций примерно соответствует представленности ученых. Из 1664 институций лишь 33% выпустили 2 и более публикации. Наиболее активно выпускавшая публикации РАН тесно соседствует с ВШЭ, тогда как следующий за ними по числу публикаций вуз, РУДН, имеет на 77% меньше публикаций, чем в среднем выпустили ВШЭ и РАН (Таблица ??). Из числа региональных ВУЗов лишь УрФУ им. Ельцина попал в топ списка организаций. Также обратим внимание на то, что в топе присутствует ЕУСПб, который по своим размерам значительно уступает всем остальным.

В разрезе цитирований из топа пропадают УрФУ, МГИМО и РУДН. ВШЭ поднимается на первое место, опережая РАН на 30%. ЕУСПб также практически вплотную соседствует с МГУ и РГГУ в середине списка, а СПбГУ соперничает с Бременским университетом Якобса.

Наиболее сильными академическими связями обладает РАН. На 40% меньшую силу связей имеет ВШЭ, остальные близко находящиеся к ним организации (МГУ, РУДН, СПбГУ, РАНХиГС) имеют примерно на 80% менее сильные связи. Обращает на себя внимание то, что в топе присутствует два чеченских университета (ЧГУ и ЧГПУ), причем один из них (ЧГУ) имеет более сильные связи, чем РГГУ и УрФУ. В рамках институционального ландшафта коллаборации ЧГУ и ЧГПУ хронологически являются довольно молодыми.

Table 24: Топ-10 по публикационной продуктивности коллабораций

НПП	Публикации,		Цитирования,		Общая сила связей	
	Организация	шт.	Организация	шт.	Организация	связей
1	РАН	1943	ВШЭ	3653	РАН	404
2	ВШЭ	1081	РАН	2563	ВШЭ	247
3	РУДН	354	СПбГУ	468	МГУ	81
4	СПбГУ	337	Бременский университет Якобса	412	РУДН	73
5	МГУ	302	МГУ	330	СПбГУ	63
6	УрФУ	164	ЕУСПб	306	РАНХиГС	61

НПП	Организация	Публикации,		Цитирования,		Общая сила связей
		шт.	Организация	шт.	Организация	
7	РГГУ	140	РГГУ	242	Чеченский государственный университет (ЧГУ)	55
8	РАНХиГС	128	Университет Куинс	228	РГГУ	39
9	ЕУСПб	122	Институт демографических исследований им. Макса Планка	195	Чеченский государственный педагогический университет (ЧГПУ)	30
10	МГИМО	93	РАНХиГС	187	УрФУ	27

При рассмотрении сети соавторства всего 394 автора соответствуют критерию в минимум 5 публикаций (Рис. 22). В этой сети один значительный компонент (связанный подграф, 28% выборки), а также несколько не связанных с данным компонентом более мелких. Этот компонент представляет из себя ядро сети, и далее мы разберем его более подробно.

Сети соавторства коллaborаций российских социологов за период 1992-2022 (фракционный счет, слева барьер – 5 и более публикаций, справа – 15 и более публикаций)

Figure 22: Сети соавторства коллaborаций российских социологов за период 1992-2022 (фракционный счет, слева барьер – 5 и более публикаций, справа – 15 и более публикаций)

При минимальном ограничении в 5 работ, опубликованных за 30 лет индексирования в WoS (5% выборки), основной компонент сети представляет из себя сеть с одним основным ядром и разветвленной периферией, которая либо находится близко к центру сети, либо слабо связана с ним единственным «маршрутом» (Рис. 23). В центре основного ядра присутствуют наиболее известные и цитируемые социологи (Ж.Т. Тощенко, З.Т. Голенкова, Иванов В.Н., Рукавишников В.О., Игитханян Е.Д., Горшков М.К. и др.), которые находятся друг от друга на определенном удалении и замыкают на себя слабее связанных авторов.

Сети соавторства российских социологов - наибольший связанный компонент сети при наличии у авторов минимум 5 публикаций за период 1992-2022

Figure 23: Сети соавторства российских социологов - наибольший связанный компонент сети при наличии у авторов минимум 5 публикаций за период 1992-2022

Если же мы строим сеть соавторства только для тех, кто выпустил за 30 лет как минимум 15 работ, в сеть попадают лишь 49 авторов (0,7% выборки), а основной компонент состоит лишь из 12 человек (Рис. 24). Сюда попадают социологи из ядра предыдущей сети. Ядро представляют авторы, соединяемые Ж.Т. Тощенко, у которого, опять же, самая разветвленная сеть. Однако в отличие от предыдущей сети, здесь сила связи исследователей примерно одинаково низкая, за исключением Голенковой З.Т. и Игитханян Е.Д. Они являются наиболее интенсивно кооперирующимися друг с другом социологами, к тому же З.Т.

Голенкова связывает между собой два участка данной сети. Социально-экономические исследователи из ФНИСЦ РАН и ВШЭ в данном случае находятся на периферии главного связанного компонента сети соавторов.

Сети соавторства российских социологов - наибольший связанный компонент сети с барьером отсечения в 15 публикаций за период 1992-2022

Figure 24: Сети соавторства российских социологов - наибольший связанный компонент сети с барьером отсечения в 15 публикаций за период 1992-2022

Таким образом, при анализе сети соавторства мы можем четко выделить ядро научной коллаборации, которое в свою очередь представлено отдельными центрами притяжения. Эти группы могут быть объединены либо вокруг конкретных персоналий, организаций, либо тематик исследований. Многие из выделенных центров притяжения сохраняются при отсечении менее продуктивных (в смысле международно рецензируемых в WoS публикаций) исследователей. Это говорит об отчетливой полигентричности такой сети и сниженной кооперации (опять же, исключительно в смысле соавторства) между более “плодовитыми” социологами. Отметим, что в целом количество продуктивных авторов и коллабораций является не очень большим – для сети соавторства скорее характерны индивидуальные работы, что может быть признаком специфических паттернов исследовательской работы с сфере социологии. Сеть достаточно фрагментарна и представлена относительно небольшим ядром коллаборирующих соавторов, среди которых международных участников нет.

Исследователи в целом отмечают рост публикационной активности в российской науке [482], при этом соотношение долей коллaborационной активности характеризуется перераспределением долей коллабораций – доля национальных коллабораций растет, в то время как доля международных снижается [270]. Для социологического сообщества также характерен рост публикационной продуктивности. Однако характеристикой коллабораций социологов является ориентация на внутренние, российские коллаборации или индивидуальную работу. В нашем случае встают два вопроса: (1) какое критериальное количество публикаций может демонстрировать индивидуальную научную продуктивность для представленного набора данных публикаций по социологии в WoS; (2) какое количество публикаций в соавторстве можно рассматривать как продуктивную научную коллаборацию, устойчивую во времени.

Индекс коллаборативности авторов, посчитанный на основе построенных сетей соавторства в программе Pajek, подтверждает структурную разрозненность и относительно невысокую склонность к выстраиванию коллабораций.<sup>17</sup> Индекс рассчитывается как единица минус отношение общего фракционного вклада автора в свои работы к общему количеству публикаций и показывает тенденцию автора к работе с другими авторами (Таблица ??). Полученные результаты соотносятся с анализом сетей соавторства, представленных графически на Рисунках 23 и 24, – даже высокопродуктивные авторы могут иметь низкий уровень коллаборационной активности. У автора может быть значительное количество публикаций, но он может работать индивидуально или с очень узким кругом соавторов.

<sup>17</sup>Данный индекс был рассчитан в программе Pajek, данные по количеству публикаций топ-авторов могут отличаться от данных, полученных в VOSviewer, так как файлы для Pajek создаются с помощью программы WoS2Pajek, которая использует встроенные алгоритмы статистической обработки данных. В целом количественно данные по топ-авторам отличаются несущественно, что позволяет проводить сравнения разных метрик. В таблице жирным шрифтом отмечены авторы с самым высоким индексом колаборативности.

Table 25: Индекс коллаборативности самых продуктивных соавторов

Автор	Общий клад автора	Количество публикаций	Индекс коллаборативности
1	TROTSUK_I	49,08	68
2	TOSHCHEN_Z	36,97	49
3	KRAVCHEN_S	35,50	38
4	#RADAEV_V	31,58	36
5	#YANITSKI_O	35,00	35
6	ZBOROVSK_G	24,17	35
7	LAPIN_N	26,28	33
8	<b>PUZANOVA_Z</b>	<b>16,53</b>	<b>33</b>
9	IVANOV_V	21,19	32
10	ROMANOVS_N	25,28	29
11	GORSHKOV_M	23,48	29
12	<b>GOLENKOV_Z</b>	<b>13,51</b>	<b>27</b>
13	BARSUKOV_S	22,33	25
14	LEVASHOV_V	21,21	25
15	TIKHONOV_N	21,17	25
16	<b>NARBUT_N</b>	<b>12,53</b>	<b>25</b>
17	FILIPPOV_A	20,00	22
18	SOKOLOV_M	18,33	22
19	TESLYA_A	21,00	21
20	STEPANOV_E	15,67	21

Коллaborации научных организаций (Рис. 25) однозначно показывают два центра притяжения – РАН и ВШЭ. Особенность положения РАН заключается в том, что она не только является центром в и так довольно связанном ядре сети организационной коллаборации (т.е. соединяет сильно связанные институции), но и открывает путь к этим коллаборациям со стороны слабо связанных организаций (справа сверху), которые, к тому же, практически не связаны друг с другом. Специфика ВШЭ состоит в коллаборации с иностранными вузами (например, Университетом Мичигана, Тильбургским университетом и др.). Между этих двух больших организаций находятся более мелкие, однако относительно ближе интегрированные с другими ВУЗы, РАНХиГС и Университет им. Г.В. Плеханова. Мы также можем отчетливо наблюдать географические группировки вокруг СПбГУ и довольно крупный кластер чеченских университетов, которые также соединяются с другими ВУЗами южных регионов России.

#### Сети соавторства организаций

Figure 25: Сети соавторства организаций

Динамически картина организационных коллабораций характеризуется преобладанием сначала РАН, потом ВШЭ, а затем региональных вузов в пространстве публикационных коллабораций. В начале 2010-х было характерно преобладание традиционно крупных московских организаций (РАН,

МГУ). Затем к ним (из крупных) добавились РУДН, СПбГУ, РГГУ, Плехановский университет, после 2015 г. – ВШЭ и иностранные университеты, и уже после 2017 г. РАНХиГС. Совершенно новые организации на академическом ландшафте, которые появились в районе 2020 г. и позже – это чеченские ВУЗы, хотя отдельные южные университеты начали свою активную деятельность гораздо раньше даже крупных московских организаций, обозначенных выше. В анализе публикационной активности в хронологическом аспекте также обращает внимание на себя тот факт, что РАН (Институт социологии) показывал положительную динамику роста по публикациям за весь период с 1992 года, в то время как начало роста публикаций по университетам-лидерам приходится на конец 2000-х – начало 2010-х гг.

В общей сложности выделены 63 страны, с которыми сотрудничают российские коллективы социологов, однако только 27 стран удовлетворяют требованию наличия в выборке минимум 5 публикаций. В топ-5 входят США, Германия, Великобритания, Италия, Нидерланды, но при этом 90% соавторства документов принадлежат России. Эти выделенные топ-5 стран представляют «традиционную» географию сотрудничества. Условная «новая» география сотрудничества включает Китай, Швейцарию, Австралию, Швецию, Испанию и другие страны.

При построении сети соавторства из всех стран за весь временной период (30 лет) обращает на себя внимание следующая особенность. Коллaborации по странам можно отнести к 2 категориям: двусторонние отношения (правая часть Рис. 7.10) и многосторонние колаборации (левая часть Рис. 7.10), куда относятся как раз страны «традиционной» географии. Такие многосторонние колаборации, безусловно, имеют больший потенциал охвата научного пространства, больше возможностей для привлечения новых участников колабораций и более высокую публичность.

Сети колабораций российских социологов по страновой принадлежности Анализ соавторства публикаций показал, что социологическое сообщество достаточно неоднородно, большое количество авторов не входит в ядро колабораций. Также публикационная активность авторов весьма невысокая – критерию порога в 5 и более статей в выборке (за 1992-2022 гг., все типы публикаций) соответствуют только 394 из 6765 авторов. Значительное количество авторов работает индивидуально, а имеющиеся научные колаборации ограничены устоявшимися коллективами из ведущих научно-образовательных организаций.

В общей сложности выделены 63 страны, с которыми сотрудничали российские коллективы социологов за период 1992-2022 гг. Первой особенностью международных колабораций российских социологов является декомпозиция на «традиционную» и «новую» географии сотрудничества. Вторая особенность колабораций – это количество стран-участников. С рядом стран выстраиваются только двусторонние колаборации, а с другими российскими социологами участвуют в многосторонних колаборациях.

Анализ соавторства организаций продемонстрировал модель сотрудничества «ядро-периферия», где ядро представлено колаборациями двух доминирующих организаций – Российской академии наук и Высшей школы экономики. Также данная модель сотрудничества характеризуется наличием группы традиционно представленных в колаборациях институтов в силу своей истории, репутации и географии (в основном Москва и Санкт-Петербург), а также присутствием относительно новых участников, что может отражать институциональную трансформацию научного ландшафта в связи с изменениями в национальной образовательной и исследовательской политике.

Международные колаборации российских социологов малочисленны и в основном представлены

российскими авторами с незначительным участием зарубежных ученых. Первым фактором, ограничивающим включенность в международные коллaborации, является языковой фактор (84,37 % публикаций представлены на русском языке). Вторым важным фактором является особенность выстраивания коллабораций – либо склонность к индивидуальной работе, либо сотрудничество с отечественными исследователями.

### **7.11 4.2.3 Картирование научного поля: применение VOSviewer и Biblioshiny на материалах Web of Science**

Анализ соприсутствия ключевых слов (keywords co-occurrence) позволяет картировать тематические кластеры ключевых слов публикаций – построить карты (сети) ключевых слов. Соприсутствие ключевых слов показывает, как соотносятся друг с другом библиометрические объекты (ключевые слова) на основе документов, в которых они одновременно присутствуют (соприсутствуют). Если ключевые слова не указаны автором публикации, то они могут быть присвоены журналом, базой данных или автоматически извлечены из заголовка, что позволяет обозначить тематическую направленность на основе метаданных академической работы [259]. В библиометрических исследованиях анализ соприсутствия ключевых слов является весьма популярным самостоятельным подходом, часто определяемым как картирование структуры знаний по соответствующему научному направлению [470].

В нашем случае картирование сетей соприсутствия ключевых слов производилось для отдельных научных областей из собранных данных Web of Science. Цель данного этапа, как и этапа обработки данных, в первую очередь была связана с определением наиболее удобного и наглядного отображения существующей тематической структуры разных дисциплин. Во вторую очередь мы попытались выделить содержательные категории, в которые объединяются встречающиеся в публикациях термины, проанализировать эволюцию популярности тех или иных терминов, выявить основные тематические тренды, а также оценить статус тех или иных тематик в ракурсе (бес)перспективности их разработке в текущем научном дискурсе. Обозначим заранее, что данная предварительная работа не является междисциплинарным анализом в полном смысле слова. Она проводилась в рамках дедуктивно определенных дисциплин (в частности, политологии и социологии) и не включает в себя перекрестные тематические совпадения между дисциплинами (например, такие точно есть между политологией и социологией).

С технической точки зрения картирование научного ландшафта осуществлялось с помощью программного обеспечения VOSviewer и biblioshiny (часть пакета bibliometrix на языке R). VOSviewer ([www.vosviewer.com](http://www.vosviewer.com)) – программа, разработанная ученым в Лейденском университете (Королевство Нидерланды) специально для построения библиометрических сетей. Разработчики программы предложили метод VOS (visualization of similarities) – визуализации сходств между объектами при построении библиометрических карт (сетей) на основе расстояний между этими объектами, которые отражают силу связи между элементами [400]. В нашем случае для работы в этой программе была сделана предобработка данных в Python с доработкой в ручном режиме, подробно освещенные в третьей части отчета. Корректировки уникальных имен (ФИО) ученых, а также названий организаций вносились через тезаурусы (справки с правилами замены встречающихся имен в метаданных на пользовательские). С помощью VOSviewer удалось произвести качественные визуализации сетей соприсутствия ключевых

слов, а также выделить кластеры терминов, тематически связанных друг с другом.

В дополнение к VOSviewer мы также использовали программу biblioshiny из пакета bibliometrix. Эта программа, разработанная учеными Неаполитанского университета (Италия), предназначена для систематического анализа как ключевых слов, так и связей между учеными, институциями, и в целом не ограничивается одним лишь сетевым анализом [27]. С помощью biblioshiny нам удалось произвести общий дескриптивный анализ публикационной активности российских ученых и университетов, а также осуществить довольно подробное первичное картографирование научного ландшафта как в разрезе «иерархии» тех или иных тем (популярность-нишевость), так и проследить изменение статуса данных тематик во времени.

Для начала представим самые общие сведения о тематиках исследований. С помощью VOSviewer были составлены топ-10 самых часто упоминаемых и наиболее важных с точки зрения соприсутствия ключевых слов. Данные списки не совпадают до конца, поскольку одного лишь упоминания в разделе «ключевых слов» недостаточно, чтобы можно было считать ту или иную тематику популярной – соответствующее ключевое слово должно не просто встречаться в большом числе публикаций, но и активно присутствовать во взаимосвязи со многими другими ключевыми словами. Чем больше терминов, с которыми соприсутствует определенное ключевое слово, и чем чаще оно с ними соприсутствует, тем более достоверно можно говорить, что данная тематика встроена в активный поток научных разработок, причем во многих областях. Чтобы учесть такое положение ключевых слов, в VOSviewer применяется показатель «общей силы связи» (total link strength). Мы ранжировали наши ключевые слова как по нему, так и по «сырой» встречаемости.

В таблицах ниже представлены топ-10 ключевых слов из социологии (табл. 1.1) и политологии (табл. 1.2). Ключевые слова выбирались таким образом, чтобы они присутствовали минимум в 15 публикациях – это позволило исключить из анализа редко упоминаемые темы, а также различные варианты написания одного и того же слова. Можно заметить, что термины, которые часто упоминаются, также имеют сопоставимо высокую общую силу связи, однако это верно не для всех ключевых слов. Хотя при ранжировке новых слов не добавляется, можно сказать, что исходя из общей частоты соприсутствия, основные тематики исследований сконцентрированы скорее вокруг молодежи и образования, тогда как исходя из общей силы связей ценности являются более популярной темой, которая широко и интенсивно присутствует в сети.

Таблица 1.1 Топ-10 терминов в анализе соприсутствия ключевых слов (социология)

Ключевое НПП слово	Соприсутствие в наборе данных (количество раз)	Ключевое слово	Общая сила связей
1 russia	178	russia	185
2 youth	93	values	119
3 education	78	youth	115
4 sociology	78	identity	110
5 identity	77	education	108
6 values	72	culture	106
7 culture	71	gender	97
8 trust	61	trust	82

НПП	Ключевое слово	Соприсутствие в наборе данных (количество раз)	Ключевое слово	Общая сила связей
9	gender	58	society	82
10	migration	58	inequality	80

Похожая ситуация, но в меньшем масштабе характерна и для политологии. Полное совпадение относительно 7 приоритетных тем – «Russia», «democracy», «China», «politics», «state», «elections», «international relations» – сочетается с некоторой вариацией популярности тем «identity», «authoritarianism» и «power», которая, однако, несущественна.

Таблица 1.2 Топ-10 терминов в анализе ключевых слов (политология)

НПП	Ключевое слово	Соприсутствие в наборе данных (количество раз)	Ключевое слово	Общая сила связей
1	russia	385	russia	326
2	democracy	142	democracy	129
3	china	139	china	113
4	politics	103	politics	88
5	state	100	state	81
6	elections	88	elections	74
7	power	68	identity	56
8	identity	67	authoritarianism	54
9	authoritarianism	58	power	54
10	international	54	international	44
	relations		relations	

На рисунках 1.1 и 1.2 представлены визуализации сетей соприсутствия ключевых слов с разбиением на кластеры для социологии и политологии соответственно. Приведем нашу интерпретацию тематических кластеров в социологии:

- *Историко-теоретический* (красный кластер): capitalism, evolution, revolution, ideology, crisis, corruption, democracy, sociology (в различных вариациях), discourse, politics, economy, society, state, Russia, China, Max Weber и т.д.
- *Социально-демографическая\_ политика* (зеленый кластер): age, children, health, family, gender, life, women, а также inequality, justice, solidarity, Europe и т.д.
- *Социальные технологии* (синий кластер): mobility, risk, behavior, identity, culture, management, modernization, globalization, innovation, integration, adaptation, tolerance и т.д.
- *Социально-экономический* (желтый кластер): labor, labor market, social structure, work, education, higher education, employment, human capital, precariat, patriotism и т.д.
- *Цифровые технологии* (фиолетовый кластер): internet, (social) media, digitalization, communication, information society, а также social networks, social capital, impact, consumption, civil society, social inequality, trust и т.д.

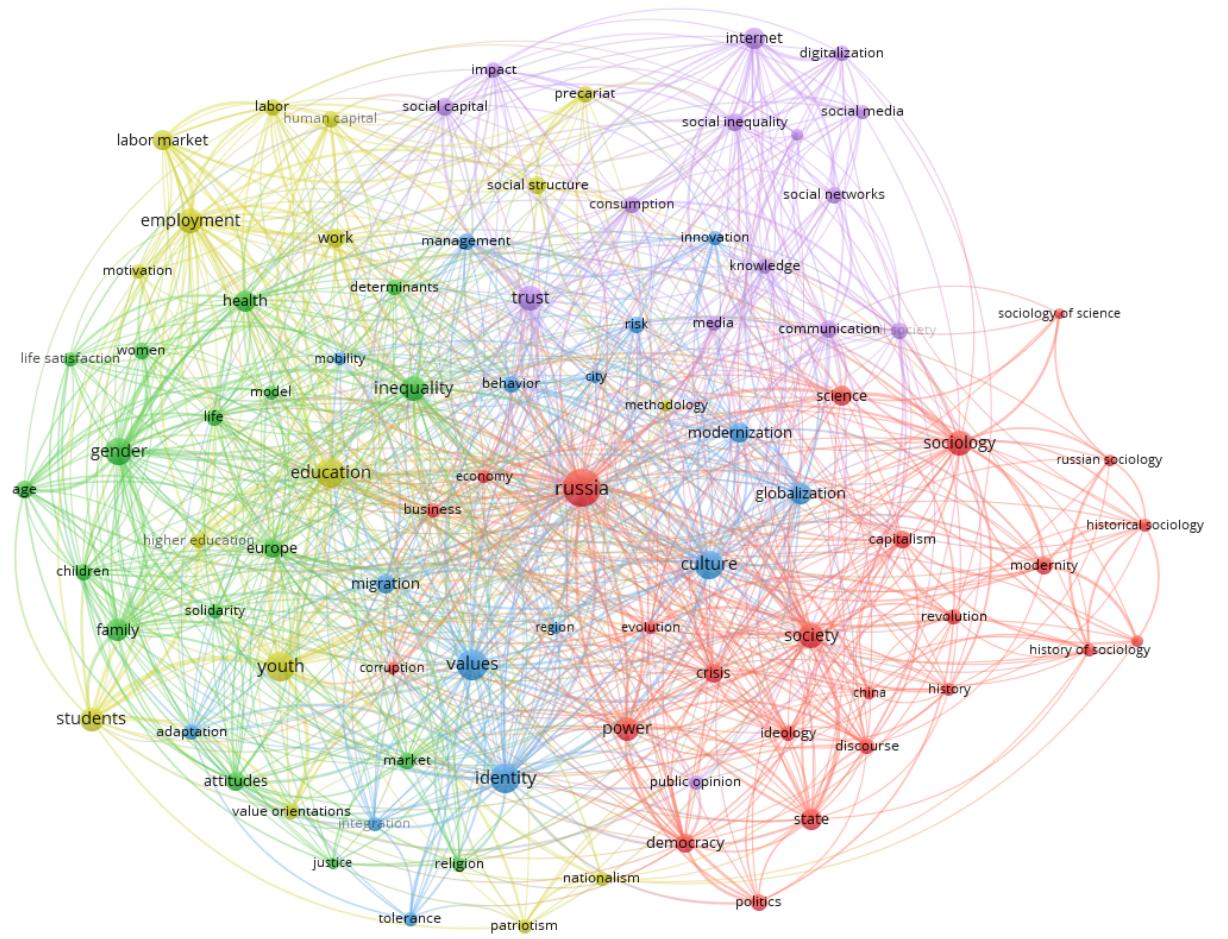


Рис. 1.1. Сеть соприсутствия ключевых слов (социология)

Теперь приведем нашу интерпретацию получившихся тематических кластеров в области политологии:

- *Геополитика и международные отношения* (зеленый кластер): Russia, China, European Union, globalization, geopolitics, regionalism, international relations, BRICS и т.д.
- *Фундаментальные политические явления* (красный кластер): politics, democracy, authoritarianism, institutions, power, governance, corruption, parties, protest и т.д.
- *Социальные институты и процессы* (синий кластер): civil society, state, ideology, values, education, modernization, law, culture, society, cooperation, migration и т.д.
- *Область социально-политической напряженности* (желтый кластер): nationalism, terrorism, war, conflict, Ukraine, crisis, identity, ethnicity, islam, trust, media, public opinion и т.д.
- *Политологический инструментарий* (серый кластер): comparative analysis, methodology, elites.

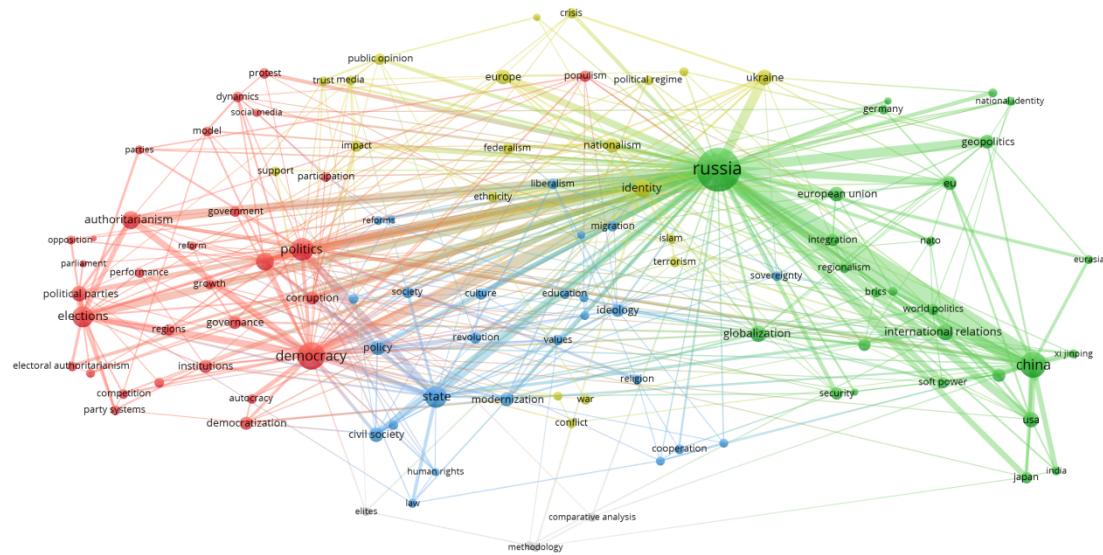


Рис. 1.2. Сеть соприсутствия ключевых слов (политология)

Представленные визуализации позволяют составить первое впечатление о том, в какой контекст встроена та или иная тема, а также увидеть потенциальные области, связки между которыми пока еще не проработаны в литературе. Например, в случае публикаций по социологии можно увидеть сравнительное отсутствие связей между темами из социально-экономического (желтого) и историко-теоретического (красного кластера), за исключением тем «России» и «образования». Или в случае политологии, достаточно заметно разграничение между областью фундаментальных политических явлений (красный кластер) и геополитикой (зеленый кластер). В случае политологии эти области, помимо России как объекта исследования, связываются через области социально-политической напряженности (желтый кластер) и, в несколько меньшей степени, через изучение социальных институтов и процессов (синий кластер). Для социологии, в данном случае, сравнительно труднее выявить связующие области и разъединенные области, однако сама идея использования сетей соприсутствия для составления впечатления о состоянии той или иной области, согласно результатам нашего анализа, выглядит продуктивной.

Анализ тематических трендов мы выполнили в нескольких видах. Во-первых, это такой же дескриптивный анализ частотности употребления тех или иных терминов (в нашем случае, заголовков, в силу малонаполненности области ключевых слов). Во-вторых, мы обратились к параметрам сетевой центральности и степени для тематических кластеров (так же составленных из заголовков), чтобы количественно оценить степень популярности и «укоренненности» тематик публикаций. В-третьих, мы выполнили анализ центральности-степени для публикаций из разных хронологических периодов отдельно, а также визуализировали общую схему эволюции публикационной активности по тематикам (также выделенных с помощью кластерного анализа).

Описание трендов по частотности тех или иных слов в заголовках можно провести как с хронологической точки зрения, так и с точки зрения длительности обращения к той или иной тематике. Так, на примере публикаций по социологии (рис. 2.1) можно выделить (хотя и с оговорками) некоторые тренды конкретного временного периода: например, появление «пандемии» и «ковида» в публикациях

2020 г., «этнического» в публикациях 2010 г. на волне беспорядков на Манежной площади в Москве, «федерализм» в связи с переустройством федерального устройства России в 2000 г. и т.д. Однако в общем виде нельзя проследить однозначного тематического тренда, т.к. практически все представленные слова из заголовков упоминаются в публикациях практически за все годы, являются долгоиграющими.

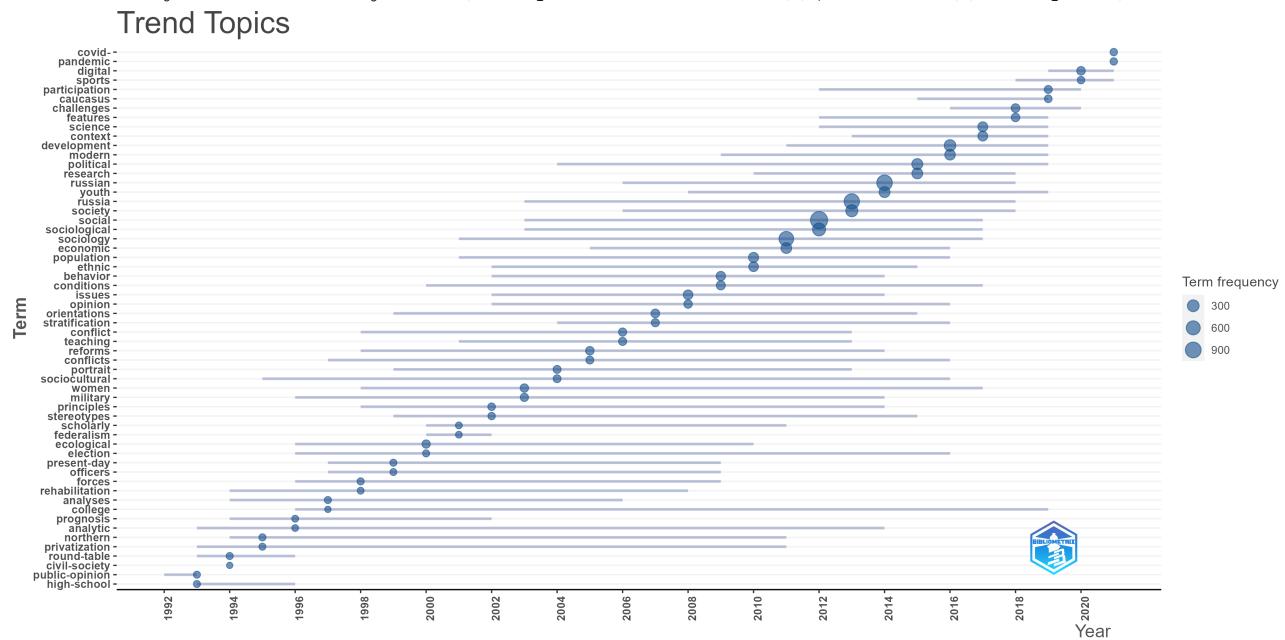


Рис. 2.1 Топ-2 популярных заголовка за каждый год (социология)

Ситуация с трендами в политологии заметно отличается от того, что происходило в социологии. Отчетливо заметно, что порядок частотности слов в целом меньше, чем в социологии, равно как и общее число выделенных терминов, несмотря на то, что для анализа отбирались 3 (а не 2, как для социологии) самых популярных слова из заголовков за каждый год. Отметим также, что использование большей части слов ограничивается серединой 2000-х годов – практически нет примеров появления одной и той же темы, начиная с 1990-х, вплоть до текущего момента. Общая тематическая эволюция показывает, что в 1990-х – 2000-х фокус внимания был сосредоточен на рыночных реформах, глобализации и модернизации России. С 2013 г. отчетливо появляется тренд на национализацию исследовательских тематик (пик частотности отдельных слов максимальен в 2014 г.). После этого можно отследить расширение «географического» фокуса в исследованиях, а также фокус на конкретные сферы практической политической деятельности и смежных сферах. Как и в социологии, наиболее популярные слова из заголовков 2020 г. касаются пандемии и ковида.

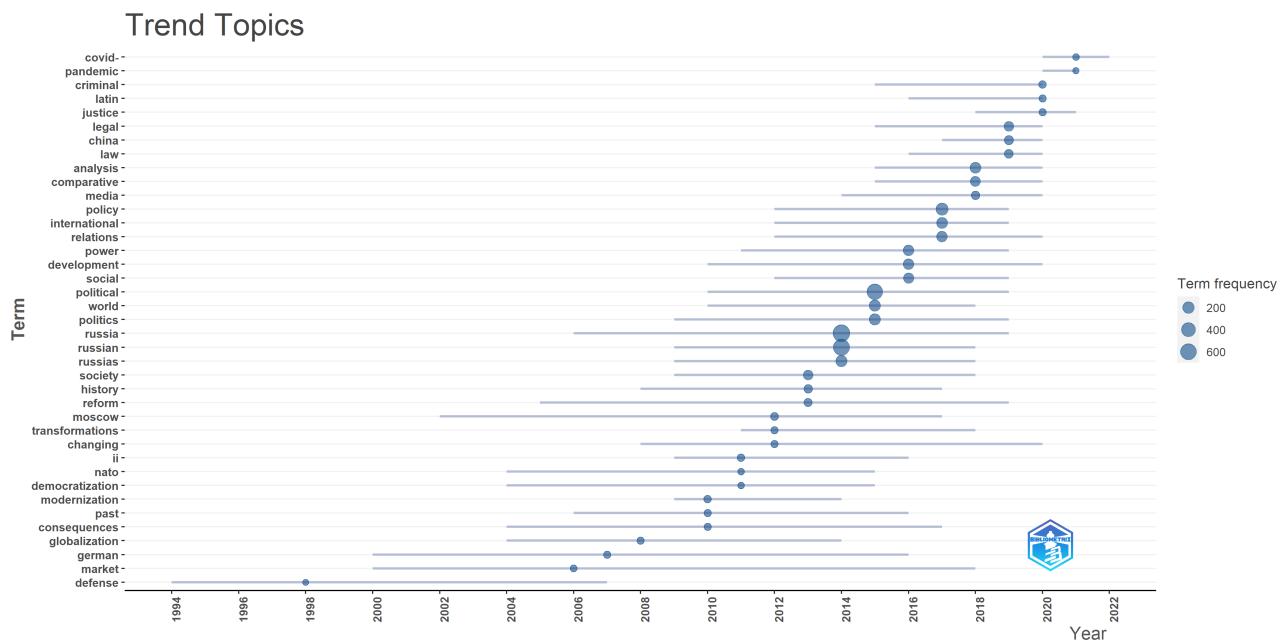


Рис. 2.2 Топ-3 популярных заголовка за каждый год (политология)

Выводы, полученные с помощью анализа частотности упоминания тех или иных слов в заголовках публикаций, являются практическим инструментом для анализа наиболее общих тематических паттернов библиографических записей. Такой анализ дает возможность получить общее представление о языке дисциплины, а также выделить некоторые исторические тренды, связанные с ее развитием. Например, в нашем случае для социологических публикаций характерно продолжительная встречаемость определенных слов на всем протяжении анализируемого периода, что может говорить о систематической роли этих понятий в языке науки. Напротив, в политологии как вариация терминов, так и их встречаемость во времени более ограничены, термины чаще отражают конкретные образования/процессы/акторов, нежели фундаментальные понятия, что также может говорить о специфике развития российской политической науки.

Тем не менее, стоит крайне осторожно относиться к этим выводам ввиду того, что встречаемость слова в заголовках – не прямой результат мотивированных действий авторов, агрегируя которые можно получить общее впечатление о мнениях и вопросах ученых, которые они озвучивают в публикациях. Для подлинно тематического анализа в идеале стоит обращаться к ключевым словам, потому что именно через них авторы определяют смысл своей публикации. Тем не менее, в наших условиях мы не могли провести анализ частотности ключевых слов, поскольку упоминания о них отсутствовали более чем в 50% библиографических записей (как политологических, так и социологических). Причины данного обстоятельства видятся в не проработанности базы данных Web of Science, однако более точный анализ может показать иные результаты.

Далее представим результаты тематической эволюции в публикациях российских социологов (рис. 3.1) и политологов (рис. 3.2). Для проведения данного вида анализа также применялся кластерный анализ, который сгруппировал публикации с тематически схожими заголовками. Темпоральные изменения в кластерах определялись с помощью других сетевых алгоритмов [78]. Временные срезы были заданы исходя из динамики публикационной активности в соответствующих дисциплинах как ориентира для общей динамики развития дисциплины.

В российских социологических публикациях стабильно выделилось меньше связанных

тематических групп, чем в политологии. Исторически тематический фокус в социологических публикациях сначала задавался политической сферой (реформы, федерализация и т.д.), затем перешел в область прикладных исследований социальной сферы (название кластера «жизнь»), а также глобальной динамики. После 2014 в исследованиях стабильно присутствовал кластер национально-ориентированных тематик, а также публикаций, сконцентрированных на фундаментальных социологических тематиках.

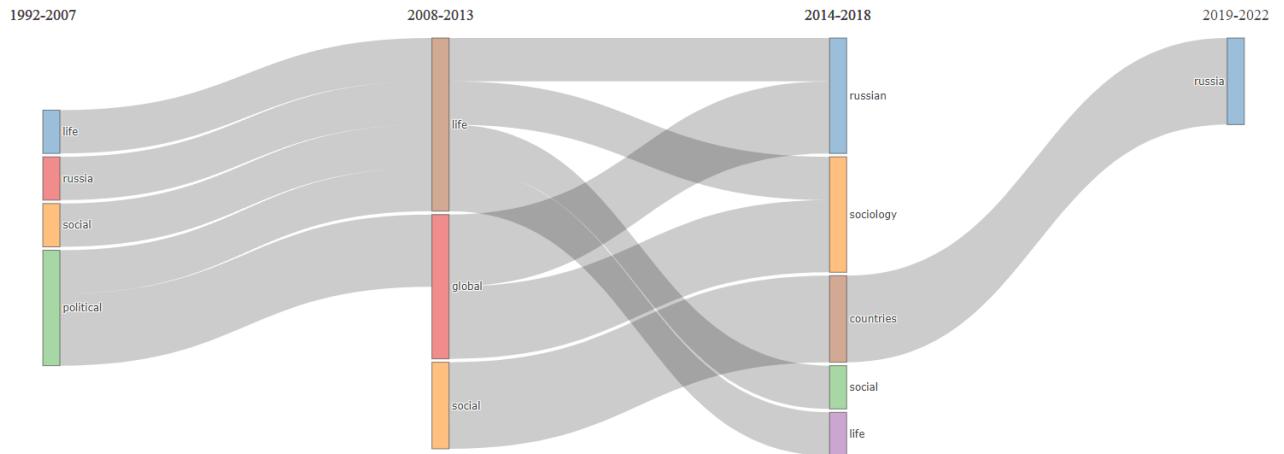


Рис. 3.1 Тематическая эволюция (социология)

В политологии вектор тематической эволюции задали работы по геополитике, а также советскому прошлому и текущим (на тот момент) конфликтам. Период 2005-2013 характеризуется широким разнообразием тематических направлений, начиная от узко-региональных (например, «Кавказ») и заканчивая фундаментальными вопросами политологической теории и политической практики. Интересно, что после 2014 г. многие темы смещаются в электоральную (или, вероятно, прикладную) область, а некоторые из национально-ориентированных тем перетекают в общую категорию «идентичности». Период с 2019 по 2022 гг. характеризуется преобладанием изучения России, а также законодательной сферы (как в России, так и в других странах).

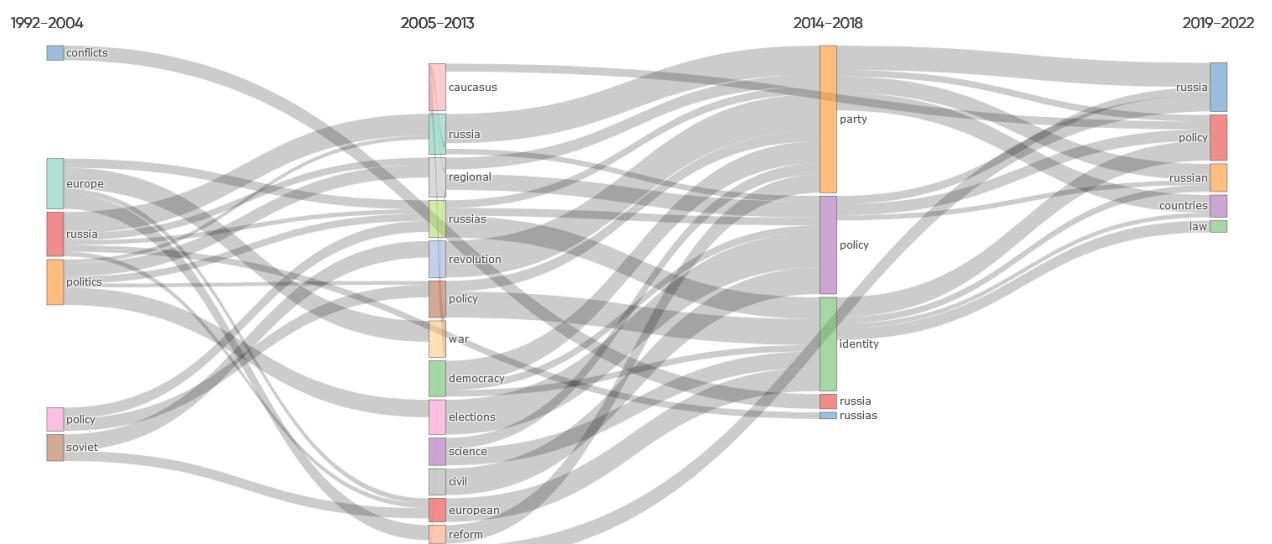


Рис. 3.2 Тематическая эволюция (политология)

Анализ тематической эволюции, в отличие от сравнительно менее изощренного дескриптивного анализа частотных трендов позволяет более глубоко оценить, какие темы и как именно существовали/трансформирова-

на протяжении времени. Благодаря использованию методов кластеризации и других сетевых алгоритмов обнаружения схожести между узлами сети, возможно построить схему преемственности публикаций, и, как следствие, отдельных тематик и тематических направлений. Более того, реализация данного подхода в biblioshiny позволяет получить доступ к полностью размеченному массиву данных, который затем можно анализировать с помощью более точных инструментов. Однако следует оговориться, что, несмотря на группировку схожих публикаций, эти кластеры остаются сугубо эвристическими, а потому нельзя делать окончательных содержательных выводов о дисциплинарной эволюции, не проведя критический анализ интерпретации алгоритмически генерированных кластерных решений.

Наконец, представим результаты структурного анализа центральности-степени тематик публикаций российских социологов (рис. 4.1) и политологов (рис. 4.2). Принципы построения тематических кластеров аналогичны описанным выше. Главное отличие – в расположении групп на двумерной оси координат, где ось абсцисс представляет степень «релевантности» той или иной тематики для остального научного поля (замеряется с помощью сетевой центральности), а ось ординат показывает степень «разработанности» тематики – количества работ в определенной области (операционализируется через сетевую степень).

Анализ схемы центральность-степень для социологии показывает, что на начальном этапе наиболее активно развивающимися и релевантными были публикации, касающиеся социологии как науки, а также студентов, образования и других социальных сфер и групп и культуры (в том числе ценностей). На следующем этапе данные темы в определенной степени укоренились, либо пролиферировались и частично перешли в область «нишевых» тем, с более узким фокусом и меньшей релевантностью для остального поля науки. Также с 2008 по 2018 гг. можно наблюдать появление области исследований, сконцентрированных на изучении глобализации, урбанизма и благополучия (они оставались «нишевыми»). В последний анализируемый период изучение образования встроилось в контекст общего социально-экономического развития, а также к этим темам добавилось изучение цифровых технологий и их последствий.



Рис. 4.1. Центральность-степень тематических кластеров во времени (социология)

Схожий анализ для сферы политологии показал следующие результаты. Изначальным двигателем публикационной активности были темы, посвященные внутренним конфликтам, а также политическим и региональным трансформационным процессам. На следующем этапе к уже имеющимся группам тем добавляется много новых, а в качестве фундаментальных закрепляются изучение демократии, пост-советского пространства и региональной политики. Проявляются работы по дискурс-анализу и правам человека. После 2014 г. тематический спектр сильно сужается, на передний план выходят темы внутрероссийской и глобальной политики, появляется небольшое, относительно незаметное число публикаций касательно украинского кризиса и электоральных систем. На последнем этапе фундаментальной темой обозначается изучение законодательной сферы России, а локомотивом выступает изучение внутрероссийской политической трансформации. Глобальный сравнительный анализ сильно уступает как в плане проработанности, так и в плане релевантности для остального дисциплинарного контекста.

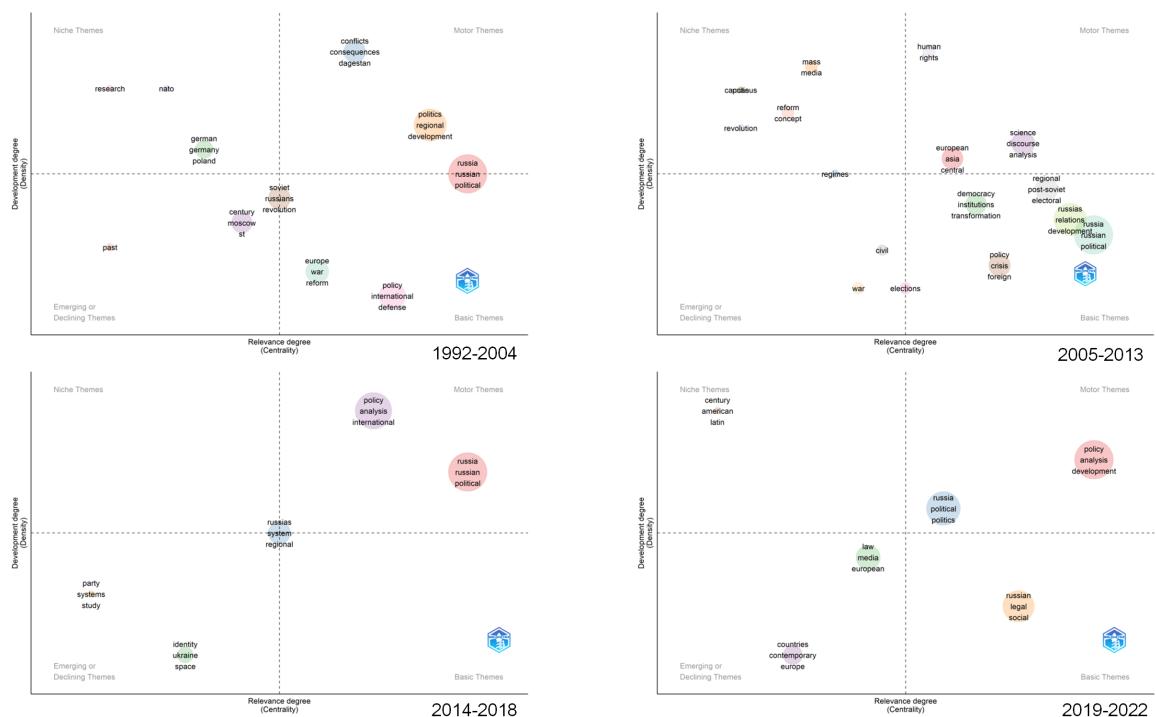


Рис. 4.2. Центральность-степень тематических кластеров во времени (политология)

В заключение отметим, что ранжирование тематик по степени их проработанности и релевантности – важный этап библиографического сетевого анализа. В отличие от предыдущих инструментов, этот позволяет увидеть структуру научного поля и понять, какое место занимает изучение того или иного вопроса, чего нельзя достичь сугубо сетевой визуализацией либо подсчетом частот тех или иных ключевых слов. Однако опять же встает вопрос содержательной консистентности результатов, полученных алгоритмическим путем, в связи с чем и к выводам данного анализа стоит относиться критически.

## **7.12 5.11.1 Модели управления благотворительными фондами – бенефициарными собственниками бизнес-компаний**

### **7.12.1 Введение**

Как мы себе представляем, что такое благотворительный фонд? Какой образ возникает в нашем воображении? Как правило возникает более-менее стандартный набор: «острая социальная проблема», самоотверженные, активные и мобилизованные проблемой люди и постоянный поиск денег. Соответственно выстраивается образ человека, его личных и профессиональных качеств – сотрудника или руководителя благотворительного фонда или НКО.

А если задуматься о другой модели, когда благотворительный фонд владеет бизнесом или группой компаний? В этой модели он сам получает доходы и сам распоряжается ими. Кроме того, в зоне его ответственности оказываются очень разноплановые задачи – не только развивать благотворительные проекты, но также и заниматься устойчивостью бизнеса и его стратегическим развитием.

В европейских странах такая модель – фонды-собственники компаний, ФСК – оказывается довольно распространенной. В числе фондов-лидеров по размеру капитала — Novo Nordisk Foundation (Дания, группа компаний Novo), INGKA Foundation (Нидерланды, группа компаний INGKA/IKEA), Fondazione Cariplo (Италия, финансовая группа Intesa Sanpaolo).

Именно эта практика и стала предметом исследования, которое мы будем обсуждать.

Один из важнейших вопросов – кто и как управляет такими фондами? Какими компетенциями обладают люди, входящие в их руководство, какие карьерные треки ими реализуются? По каким принципам формируются команды руководителей? Поэтому мы будем обсуждать не только уже полученные результаты работы. Мы попробуем определить принципы следующего исследования и наметить шаги, направленные на изучение профессиональных сетей и связей руководителей таких фондов, их профессиональный бэкграунд и «композицию» советов директоров, необходимую для успешной реализации непростых и довольно разноплановых задач развития.

В российской практике такая модель управления практически не представлена. Только в последний год наметилось движение в эту сторону у некоторых фондов и компаний. Вместе с тем, определенное развитие получила смежная форма управления и финансирования социальных и благотворительных программ – эндаументы, фонды целевых капиталов. Их главное отличие от «industrial foundations» в том, что сфера деятельности эндаументов ограничена финансовым капиталом, тогда как ФСК также владеют акциями компаний, участвуют в получении дивидендов и в той или иной степени и объеме участвуют в управлении компаниями.

### **7.12.2 Цели и задачи исследования**

Цель работы — сформировать углубленное представление о моделях управления ФСК в странах, где подобные практики широко применяются и имеют существенный вес в экономике, а также определить возможности и стратегии для развития рассматриваемых сценариев в российских реалиях.

Для достижения цели в исследовании формируются следующие задачи: - изучить, что является мотиватором собственника при принятии решения о выстраивании управления активами через создание фонда; - сформировать представление о том, что представляет собой структура управления фондом и как оценивается ее эффективность; - проанализировать, как выстроено управление с бизнес-активами и

взаимодействие с другими значимыми участниками (основателями, наследниками, советом директоров и др.), выявить экономическую обоснованность управления активами посредством фонда; - описать структуру и приоритеты благотворительных программ и программ социального развития, оценить факторы их устойчивости и эффективности, - описать страновые регуляторные особенности, влияющие на деятельность IF - определить другие преимущества такой моделью управления бизнесом, понять, для кого и как результат такого управления является или может являться полезным и выгодным.

### **7.12.3 Практическая значимость исследования**

Модель, когда благотворительный фонд выступает собственником успешного бизнеса, оказывается распространенной в Европе, в первую очередь – в странах Северной Европы и в германоязычных странах. В долгосрочной перспективе именно такие решения оказываются наиболее эффективными и демонстрируют стабильность в осуществлении социальных программ и проектов.

В России в настоящее время активно растет сектор целевых капиталов: законодательно эндаументы имеют возможность становиться собственниками больших активов. Однако в российской практике таких примеров пока не много.

В исследовании наглядно продемонстрирована важная роль страновых особенностей при выборе модели управления фондами. Практическая значимость работы – в задаче наметить направления развития в масштабах российского сектора благотворительности и филантропии. Результаты могут быть использованы для анализа достоинств и недостатков такого рода решений, сравнительного анализа структуры собственности, формирования более комплексного взгляда на инфраструктуру благотворительности, для расширения нашего понимания собственности и способов управлять ею.

По итогам работы были сформулированы рекомендации для собственников компаний, задумывающихся о стратегической устойчивости корпоративных социальных программ, для благотворительных фондов, управляющих эндаументами, для государства – по развитию регулирования в этой сфере.

### **7.12.4 Методы сбора и обработки данных**

Исследование построено в формате кейс-стади. Оно опирается на анализ и обобщение существующих практических примеров десяти фондов-собственников компаний и предлагает выводы о применимости этого подхода в российских реалиях. В каждом из кейсов рассматривается перечень вопросов, помогающих выполнить задачи исследования.

Авторами были разобраны и проанализированы десять фондов – бенефициарных владельцев компаний: Novo Nordisk, Bosh, IKEA, Карл Цейс, Hempel, Ramboll, Pierre Fabre, индийская Tata, итальянские банки Intesa Sanpaolo и La Caixa.

Тематическая структура case-study [461]: 1. История возникновения фонда, компании: основные моменты 2. Миссия и цели фонда 3. Структура владения и управления компанией 4. Владение 1. Управление 2. Процесс и ступени принятия решений в фонде относительно компании 4. Финансовые показатели фонда и компании 5. Регуляторный режим

На этом этапе методология работы полностью ориентировалась на принципы разработки «делового кейса». Однако следующий аналитический шаг нацелен на изучение практики формирования советов директоров ФСК с использованием сетевого подхода. Изучение состава СД фонда Ново Нордиск

за десять лет показало, что его члены не только перемещаются по значимым позициям внутри всей холдинговой структуры, но также включены в управляющие органы других важных стейкхолдеров. Таким образом в СД фонда аккумулируется необходимая экспертиза и социальный капитал каждого из участников. Сетевой подход представляется очень уместным для понимания сути и принципов одной из ключевых практик функционирования ФСК.

#### 7.12.5 Обзор литературы

Практика, когда БФ являются бенефициарными владельцами бизнес-активов, не развита, она только начинает формироваться. В российских исследованиях все больше появляется работ про межпоколенческий трансфер российских бизнесов и частных благотворительных проектов [468]. Много работ посвящены барьера姆, сдерживающим развитие ФЦК, эндаументов: неразвитость фондового рынка, недостатки в управлении налоговыми льготами, трудности взаимоотношений с донорами, с управляющими компаниями и другие [17] [[450]][480][[476]][471]

«Фонды – собственники компаний» (industrial foundations) – это организации, владеющие акциями или же долей одной или нескольких бизнес-образований напрямую или же через холдинговую структуру [386]. Формы ФСК зависят от специфики законодательства в конкретном государстве. Это могут быть фонды, учреждения, трасты с образованием или без образования юридического лица. Но их все объединяет параметр – у них нет собственников или акционеров.

Зачастую ФСК создается основателем и акционером компании и такой выбор является альтернативой передаче прав владения компанией наследникам или внешним инвесторам. Другой вариант их возникновения, когда бизнес изначально создается со стратегической социальной целью (как Novo Nordisc), под влиянием государственного регулирования (Фонд Карипло, Италия) или же по решению наследников (трасты Tata). В большинстве случаев это некоммерческие организации, они владеют или контрольным пакетом акций или особыми классами акций [155]. Существует практика, когда акции передаются в фонд на принципе безотзывности [155].

Модели управления, принципы формирования совета директоров, приоритеты деятельности, структура социальных и благотворительных программ, а также традиции целеполагания в значительной мере определяются страновой спецификой.

Выделяют четыре параметра вариативности форм и решений [344]: - традиции филантропии как в стране, так и в семье основателя; - социально-экономические и культурные особенности страны; - деловые процессы в компании и динамика в «ее» секторе экономики; - специфика и изменения в законодательстве.

Модель ФСК принципиально отличен от грантодающих и благотворительных НКО. Набор задач в них более широкий и требует более широкого набора компетенций и профессионального опыта. Однако главная цель ФСК – реализация акционерных прав согласно с принятыми положениями и ценностями.

В литературе и в исследованиях ФСК выделяются несколько важных для темы направлений.

Управленческая дистанция [155]. Предметом анализа становятся полномочия советов директоров, распределение функций контроля в Фонде и компании, в управлении операционной деятельностью бизнеса, состав и принципы формирования советов директоров компании и фонда, в какой мере распространено пересекающееся членство. Оценивается, насколько ФСК оказываются вовлечеными и активными собственниками. В качестве акционера фонд участвует в назначении своих представителей в

органах управления компании и участвует в утверждении основных решений, контролирует финансовое положение и др. [53] [165] [343] [58]

Исследования показывают, что грамотно выстроенное управление ФСК и корпоративного управления оказывают положительное влияние на эффективность компаний – независимо от системы финансового стимулирования директоров [194]. Тем не менее, здесь фиксируется нелинейная зависимость между управленческой дистанцией и эффективностью компании: позитивное влияние возрастающей дистанции после определенного момента демонстрирует обратный эффект [155].

Назначение директоров и работа СД. Исследования фокусируются на большом спектре вопросов и проблем: принципы формирования, оценки рисков для стратегического управления в зависимости от принятых принципов формирования СД, роль законодательного регулирования общественная подотчетность благотворительных фондов [[388]][155]. Еще одна большая область исследований - система мотивации и финансовые условия участия в СД в ФСК, анализ КРП и принципов его работы, проблематика нефинансовой мотивации и т.д. [155]

Страновая специфика. Большой сегмент исследований посвящены страновой специфике и изучению социо-культурных и социально-экономических детерминант в развитии практики ФСК. Наибольшее распространение ФСК получают в Северной Европе и германоговорящих странах, хотя они встречаются и в других странах мира и бывают в различных формах [[392]][194][344]. Проводится оценка их вклада в ВВП стран и в объем благотворительного сектора [391]. Систематический обзор страновых отличий проведен исследователем Центра корпоративного управления БШ Копенгагена С. Томсеном [389].

Синдром «сиюминутного мышления». В работах, посвященных корпоративной социальной ответственности и стратегическому развитию институциональных условий для сектора благотворительности эта проблематика обозначается концептом «short-termism»: стремясь в моменте максимизировать возможные выгоды, «временщик» не просто теряет выгоды в стратегической перспективе. Такие действия и такой стиль мышления проецируются в общество и формируют специфический тип культуры и тип социальных отношений. Исследования нацелены на анализ механизмов, которые помогают бороться с «сиюминутным мышлением» и как корпоративной практикой, и как социальным феноменом [390]. Исследователи показывают, что ФСК как специфическая форма управления и сами компании, принадлежащие благотворительным фондам, оказываются хорошим инструментом для преодоления синдрома, т.к. с ними связано стабилизирующее влияние на уровень и качество занятости, на развитие местной экономики. Кроме того, фирмы, принадлежащие фондам, лучше справляются с экономическими кризисами, показывают более высокую выживаемость в длинной временной перспективе [390]. Например, для них выше вероятность выживания в 40-летней перспективе (выживают 30% от их числа – против 10% для компаний с другой формой управления). Причем эти же исследователи показали, что такая форма управления компанией показывает свою эффективность и в кризисное время, когда требуются быстрые решения и решительные действия.

Семейный бизнес. В работах показано, что ФСК имеют много общего с практиками семейного бизнеса [390]. Однако некоторые из базовых параметров ФСК делают его еще более долгосрочным предприятием, чем только лишь семейное владение, поскольку иначе решается вопрос преемственности и собственности, вопросы управленческих и предпринимательских компетенций преемника, иное отношение к дивидендам, иные горизонты планирования и основания для построения долгосрочных

планов [[386]][[46]][[371]][[391]].

#### 7.12.6 Обсуждение

Главный вопрос и миссия нашей работы – поиск ответа на вопрос: почему нам и нашему государству необходимо интересоваться этой темой?

Ответ на этот вопрос лежит в плоскости экономического развития, драйверов устойчивости корпоративного управления, детерминант появления долгосрочных бизнес-стратегий и развития инфраструктуры благотворительности. Такая аргументация важна в контексте осмысления 30-ти летнего опыта формирования рыночной экономики в нашей стране. По мере ее усложнения и развития появляются возможности создавать все более сложные и долговременные институты.

Мы видим несколько аргументов, которые позволяют отнести к Фондам предприятий как к стратегической задаче экономического и социального развития [461]. 1. Постепенно менять отношение россиян к «богатству», к «богатым людям», и в том числе – к занятию предпринимательством [356]. ФСК как модель владения и управления оказывается противопоставленной личному владению бизнес-активами, что потенциально может сформировать иные представление и новое понимание «частной собственности». Можно предполагать, что практика коллективного принятия решений по вопросам управления компаниями (особенно крупными) будет с большим пониманием воспринято нашими согражданами, чем «эгоистичное владение» [461]. 2. Государству, чтобы реализовывать долгосрочные экономические стратегии, необходим мандат легитимности относительно базовых инструментов, с помощью которых оно осуществляет регулирование социально-экономических процессов [461]. ФСК и принадлежащие им компании со временем могут стать устойчивым и эффективным институтом, создающим долговременный эффект и не требующим непосредственного управляющего участия государственных структур в его деятельности. 3. Описываемая модель корпоративного управления снимает с государства необходимость регулировать частные капиталы и определять их судьбу – со всеми сопутствующими этой активности негативными последствиями: конфликтами элит, коррупцией, разрушением устойчиво действующих компаний, компрадорскими установками элит и т.д. [461]. Кроме того, ФСК оказываются одним из наиболее эффективных инструментов, которые заставляют деньги оставаться и работать в стране. Причем делают это не принуждением, а через создание возможностей нового типа. 4. Владение фондом, по-видимому, является примером частного предприятия другого типа, которое не подвержено недостаткам финансализации – сведения роли бизнеса только лишь к зарабатыванию денег и формированию частного капитала [108]. Владение фондом также, по-видимому, позволяет избежать проблем наследования, семейных конфликтов и кумовства, которые преследуют семейный бизнес [53]. Более того, фонды являются частными организациями, которые не подвержены знакомым проблемам государственных предприятий, таким как политическое вмешательство или мягкие бюджетные ограничения [389]. 5. Что делает владение фондами особенно перспективной формой корпоративного управления, так это сопряжение благотворительности с бизнесом [461]. Фонды должны и могут служить неэгоистичным целям и ценностям, разделяемым в обществе. В отличие от англо-американской традиции, регулирующие органы в странах Северной Европы рассматривают владение компанией как служение полезной социальной цели. Поэтому их законодательство изначально ориентировалось на создание инфраструктуры для деятельности социально-ответственных акторов

[389]. 6. Многие исследователи рассматривают Фонды предприятий как способ борьбы с «парадоксом временщика» [461]. Фонды являются долгосрочными владельцами, что может дать их компаниям, конкурентное преимущество в некоторых сферах бизнеса, например, в фирмах, интенсивно занимающихся НИОКР, с длительным жизненным циклом продукции [53]. Похоже, что преимущества долгосрочного подхода более выражены в крупных фирмах, чем в небольших стартапах.

Несколько факторов, по-видимому, усиливают практику долгосрочного планирования в компаниях, принадлежащих фондам [461]. Например, консервативная долгосрочная ориентация может быть заложена базовыми правовыми положениями. Например, в ряде стран есть положения, которые обязывают Фонды сохранять полученные активы. В отличие от компаний, они не могут (за исключением случаев банкротства) быть распущены до тех пор, пока их бизнес-актив остается действующим [[344]][[387]]

#### **7.12.7 Результаты**

Основной результат работы – систематическое описание десяти отобранных кейсов фондов – бенефициарных владельцев компаний: Novo Nordisk, Bosh, IKEA, Карл Цейс, Hempel, Ramboll, Pierre Fabre, индийская Tata, итальянские банки Intesa Sanpaolo и La Caixa. Для каждого из примеров были описаны [461]: 1. История возникновения фонда, компании: основные моменты 2. Миссия и цели фонда 3. Структура владения и управления компанией 1. Владение 2. Управление 3. Процесс и ступени принятия решений в фонде относительно компании 4. Финансовые показатели фонда и компании 5. Регуляторный режим

Кроме того, были приведены аргументы в пользу подобной практики развития корпоративного управления с точки зрения собственников и акционеров компаний, с точки зрения государства и в контексте развития социальных и экономических процессов.

### **7.13 5.11.2 Изучение гендерной специфики деструкторов руководителей крупных российских компаний**

#### **7.13.1 Введение**

В научной литературе и в эмпирических исследованиях существует множество работ, посвященных изучению гендерной специфики лидерства. Существуют «свои» традиции в самых разных дисциплинах – в социологии, политической науке, в психологии и социальной психологии. Заметная доля публикаций приходится на междисциплинарные области, такие как феминистические и гендерные исследования, изучение организаций и менеджмента, сочетающие в себе самые разные жанры: публицистика, научно-популярно изложение теорий и концепций, описание лабораторных экспериментов и полевых исследований, а также обобщающие работы, использующие методологию мета-анализа. Однако увеличение эмпирических данных не приводит к целостному пониманию лидерства как феномена и его гендерной специфики. Можно согласиться с исследователями, которые давно говорят о накопленных знаниях как о множестве несвязанных, противоречивых, фрагментарных сведений и фактов, о череде одновременно амбициозных и тривиальных теорий и исследований [341].

На фоне большого числа публикаций о лидерстве в целом и женском лидерстве – в частности, заметно меньшее число публикаций посвящено «деструктивному лидерству». Многие авторы утверждают, что этого рода публикации фрагментированы, основаны на разношерстной

методологии, их совокупность плохо осмыслена [383]. Тем не менее, на основании проведенного анализа литературы и исследований мы можем предположить, что существует гендерная специфика в проявлении деструкторов у руководителей, а ряд исследователей и подходов дают сильные объяснительные модели для интерпретации проявляющихся гендерных отличий. Именно этот сюжет стал предметом нашего исследования. Источником данных послужила база данных результатов оценки руководителей российских компаний по методике Хогана, собранная за период с 2019 по 2021 год.

Мы не можем утверждать, что имеющийся массив репрезентативен для всех российских руководителей: он формировался в результате коммерческой деятельности нескольких компаний, предоставляющих услуги по оценке руководителей. Тем не менее, для задач исследования важны другие характеристики: точное попадание в целевую группу и структура массива, позволяющая возможности сравнивать разные категории руководителей.

Массив данных, который использовался для статистической обработки, был полностью анонимизирован. У исследователей не было доступа ни к персональным данным респондентов, ни к данным о конкретных компаниях.

**7.13.1.1 Цели и задачи исследования** Научная проблема: проанализировать гендерные различия в деструкторах топ-менеджерах российских компаний, опираясь на эмпирические данные. В основе эмпирики – датасет на основе психометрического теста Хогана [171]. Проанализированная структура деструкторов должна быть соотнесена с традицией психоаналитических исследований лидерства и токсичного лидерства.

#### **Задачи исследования:**

1. Разобраться с понятием «деструкторы» в контексте проблематики «токсичного лидерства»; углубиться в специфику их измерения в методике Хогана;
2. Провести анализ имеющегося дата-сета с прицелом на гендерной специфике деструкторов; найти ключевые деструкторы, где есть дифференциация по гендерным группам; изучить возможности анализа гендерных отличий деструкторов по различным отраслям.
3. Оценить эвристические возможности психоаналитического подхода к изучению деструкторов руководителей; увидеть их связь со спецификой женского и мужского лидерства.

С точки зрения авторов исследования, именно психоаналитическая традиция системно ведет разработку природы деструкторов и рассматривает бессознательное в качестве их природы. Традиция формируется работами таких ученых, как Фрейд, Лакан, Кляйн, Баум, Клули, Кетс де Врис, Миллер, Штейн, Залезник, Костас, Тахери, Драйвер, Обхольцер, Маккоби, Кохут, Габриели, Мак-Вильямс и другие. Целый ряд направлений изучает деструкторы как связь между невротическими дисфункциями лидера и организационной патологией, взаимозависимость и взаимное влияние лидера и его последователей, различие между лидерами и менеджерами, аутентичное лидерство, опираясь на клиническую парадигму, объясняют природу деструкторов и связанное с ними иррациональное поведение.

Также можно сказать, что для психоаналитической школы больший интерес представляет влияние ранних периодов на лидерский потенциал и не гендер как таковой, а скорее роли отца и матери в развитии и формировании не только лидерской позиции, но и лидерских деструкторов. Традиционно лидерская

тема, включая гендерный аспект, рассматривается с точки зрения прохождения нарциссической стадии и формирования «нормального» или «патологического» нарциссизма у обоих полов. Можно резюмировать результаты анализа следующим образом: в вопросе формирования деструкторов лидеров основная роль принадлежит не гендеру, а специфике прохождения стадий развития и влиянию родителей. При этом, похоже существуют отдельные социальные и культурные нормы, которые определяют различия в том, как матери относятся к сыновьям и дочерям, что формирует специфику в проявлении лидерства.

**7.13.1.2 Практическая значимость исследования** Новизна работы состоит в использовании психоаналитической традиции для понимания природы гендерных отличий в деструкторах, обнаруженных у руководителей российских компаний. Фокус исследования конструируется пересечением четырех тем и подходов: лидерство, деструкторы, гендер, психоаналитический подход. Проблематика в этих направлениях «приземляется» на эмпирические данные выбранной целевой группы: реально действующие руководители российских компаний.

**Практическая значимость** работы состоит в комплексности подхода и в применимости полученных результатов для задач бизнес-консультирования, персональной работы с теми, кто претендует на лидерские позиции, для оптимизации командной работы и гармонизации внутреннего взаимодействия команд, для решение задачи управлять деструктивными характеристиками лидера (лидеров).

### 7.13.2 Обзор литературы

«Темная сторона» лидерства была предметом научного изучения, по крайней мере, в последнее десятилетие [340]. Исследователи приписывают этой теме такие термины, как деструктивное лидерство [109], нарциссическое лидерство [329] или токсичное лидерство [239], и это лишь некоторые из них.

Первое впечатление, которое ожидает исследователя темы деструкторов и «темной стороны» лидерства, связано с отсутствием общепринятого определения данного понятия.

Так Хиггс [164] указывает, что в настоящее время не существует общепринятых определений и разграничений для термина «деструктивное лидерство». Уместнее говорить о том, что «деструктивное лидерство» является зонтичным понятием. Также Хиггс утверждает, что ключевыми словами, которые определяют этот термин, являются «дирейлеры лидерства», «токсичное лидерство», «лидерство темной стороны», « злоупотребление лидерством» и «деструктивное лидерство». Аналогичным образом, Теппер [383] утверждает, что литература по данной тематике «фрагментирована, плохо интегрирована и использует различные методологии». С этой точки зрения интересны работы, структурирующие данное понятие. Так исследование Чжан, Лесли и Ханну [440] помогло определить несколько слабых областей лидерства, обобщенных позднее Коутом [87]. По данному исследованию деструкторы лидера с поведенческой точки зрения можно определить как проблемы с производительностью, отношениями, изменениями, созданием и управлением командами, опытом.

Липман-Блюмен [239] относят токсичное лидерство к процессу, в котором лидеры в силу своего разрушительного поведения и/или дисфункциональных личностных характеристик наносят серьезный и устойчивый вред своим последователям, организациям и другим людям. Автор делает упор на политических лидеров, хотя его выводы распространяются и на организационный контекст. Он так же, как и Теппер, включает в свой подход разных акторов (лидер, подчиненный, последователь и

т.д.) и влияние контекста, истории, культуры. Он указывает, что определение токсичных лидеров - нетривиальный процесс, так как токсичный лидер для одного человека может считаться героем для другого.

Липман-Блюман один из немногих авторов, кто разделяет поведенческие и личностные характеристики токсичной или «темной стороны» лидерства, а также обращает внимание на разнообразие токсичного лидерства из-за различной степени его проявления и уровня осознанности при демонстрации такого поведения. Чтобы проработать концепт токсичного лидерства, Липман-Блюмен предлагает использовать многомерный фреймворк: учитывать интенциональность или целенаправленность поведения токсичных лидеров, уровень интенсивности их токсичности, типы деструктивного поведения, типы дисфункциональных личностных качеств, которые управляют их решениями и действиями, и последствия самих решений и действий.

Можно обратить внимание, что деструкторы лидерства чаще всего изучаются с поведенческой точки зрения, через описание того, как ведут себя токсичные руководители и какое влияние они оказывают. Однако нас в первую очередь интересует не только проявление деструкторов, но их природа. Многие авторы описывают деструкторы как иррациональное поведение, иногда не поддающееся объяснению ([383], [31], [109], [239]). Значительная часть нашей мотивации и поведения происходит вне осознания, и мы не имеем полного контроля над нашими процессами восприятия ([195], [197]).

Считается, что большая часть психоаналитических подходов к изучению деструкторов опирается на концепции Фрейда и Кляйн [92]. К авторам этого направления принято относить Баума, Клули, Кетс де Бриса, Миллера, Штейна, Залезника, а меньшее количество публикаций также используют теорию Лакана [92]. Например, Костас и Тахери, Драйвер. Психоаналитический взгляд на лидерство основан на том, что невозможно иметь представление о вопросах развития авторитета и лидерства без учета внутреннего мира индивида [292]. При этом сам предмет изучения лидерства в психоаналитическом подходе достаточно комплексный и можно выделить следующие направления в области исследований деструкторов лидеров:

1. Исследования, раскрывающие связь между невротическими патологиями, такими как нарциссизм, и организационными дисфункциями ([197], [196]; [254]; [329]; [345]).
2. Исследования лидерства, фокусирующиеся на вопросе связи лидера и последователей. Так Кохут [463] популяризовал идею о том, что последователи часто рассматривают лидеров и относятся к ним так, как это отражает их более ранние детские отношения с родительскими фигурами. Габриэль [136] изучал различные фантазии с помощью историй, которые последователи рассказывают о своих лидерах. Некоторые также изучали эмоциональную связь между лидерами и последователями ([195], [197]). Кетс де Брис показывает, как эта связь между лидером и последователями часто включает в себя защитные механизмы, процесс переноса, а также управление тревогами, что в комплексе может приводить к дисфункциональному поведению.
3. Исследования, фокусирующиеся на различиях между менеджерами и лидерами. Например, подход Залезника [196], в котором он выступает за развитие лидеров, а не менеджеров в организациях. Аргументируя это тем, что в то время, как менеджеры сосредоточены на рациональном, порядке и контроле, лидеры более интуитивны, эмоциональны и креативны, и у них есть страсть и видение. Лидеры более эмоциональны и могут лучше справляться с неопределенностью. Именно менеджеры,

по мнению Залезника, ставят процесс, политику и личные интересы выше содержания работы [196].

Что, в свою очередь, делает их очень манипулятивными, хитрыми и расчетливыми.

4. Исследования, которые опираются на подход Лакана. Драйвер [107], фокусируется на изучении лидерских идентичностей, которые в конечном итоге приводят к неудаче. Костас и Тахери [86] обсуждают популярный в настоящее время подход «аутентичного лидерства» и его акцент на положительных эмоциях.

Поскольку понятие деструкторов или «темных сторон» лидерства оказывается сложносоставным, комплексным, всегда существует потребность в некоторой систематизации и синтезе этих понятий для практического использования. В психоаналитической школе таким исследователем можно считать Мак-Вильямс, которая предложила выделить девять типов организации личности, опираясь на два измерения: определение личностной структуры, которое основано на возрастном периоде переживания травмирующего события, и сочетание защитных механизмов. Объединяя указанные выше понятия, Мак-Вильямс определила типы организации личности или треки, изучая и диагностируя которые, можно определить стиль лидерства и специфическое проявление его «темной стороны» или деструктора.

Хорошим примером практической реализации теоретических подходов оказываются психометрические инструменты. В частности – это методика Хогана и ее возможности для оценки деструкторов руководителей. По мнению Хогана, — это не просто отсутствие навыков, а скорее, дисфункциональные наклонности и связанное с ними поведение [171]. Специфика таких личностных характеристик или деструкторов заключается в том, что они могут ухудшить или нейтрализовать любые навыки и компетенции, которые принято относить к «светлой стороне» лидерства. Хоган указывает, что феномен «дисфункциональных диспозиций» характеризуется сочетанием технической компетентности и межличностной неадекватности. В худшем случае лидеры, которые транслируют деструкторы, могут восприниматься своими подчиненными и организацией в целом как «абьюзивные» [383] или «тираничные» [31]. Так мы видим, что подход Хогана сочетается с другими моделями деструктивного лидерства и скорее представляет собой изучение причин в виде дисфункциональных личностных характеристик или деструкторов, которые на уровне

поведения приводят к негативным проявлениям и последствиям.

Хоган считает себя приверженцем социоаналитической теории [169], которая, по его мнению, отражает динамику, связанную с успехом в любом групповом процессе. Социоаналитическая теория утверждает, что, как животные, живущие в группах, люди разработали стратегии для максимального индивидуального и группового выживания [169]. Все группы организованы по статусной иерархии, и взаимодействие внутри групп включает в себя две темы: о принятии и о статусе. Соответственно, люди мотивированы как “ладить с другими” (максимизировать популярность), так и “продвигаться вперед” (максимизировать статус по отношению к другим членам группы).

Такой подход предполагает, что организационный контекст тесно связан с социальным взаимодействием и требует реализации этих мотивов: ладить с другими и продвигаться вперед [170]. Чтобы реализовать первый мотив – «ладить с другими» - люди должны сотрудничать, а также восприниматься дружелюбными и позитивными. Чтобы «продвинуться вперед», необходимо проявлять инициативу, стремиться к ответственности и стараться быть признанными.

Природу деструкторов как дисфункциональных наклонностей Хоган раскрывает через теории

Адлер, Хорни, Эриксон, Салливан [171], которые утверждали, что проблемы людей основаны на том, как они взаимодействуют с другими. При этом, ранний опыт (в семье, в школе, в группе сверстников) гарантирует, что почти каждый чувствует себя неадекватным в отдельных аспектах. То есть детство и юность почти неизбежно сопряжены со стрессом, и у большинства людей развиваются ожидания, что в определенных ситуациях их будут критиковать и/или они будут чувствовать себя неадекватными или беспомощными.

Ключевой теорией для определения природы деструкторов Хоган считает работу Хорни, которая выделила 10 «невротических потребностей». Позже она обобщила эти потребности в терминах трех тем [483]:

1. Движение навстречу людям — т.е. управление своей неуверенностью путем создания союзов, в которых угроза критики может быть сведена к минимуму.
2. Отдаление от людей — т.е. управление своим чувством неадекватности, избегая истинной связи с другими.
3. Движение против людей — т.е. управление сомнениями в себе, через доминирование и запугивание других.

Опираясь на приведенные выше понятия и теории, в 1997 году Хоган разработал опросник The Hogan Development Survey ([171]) для измерения и оценки вероятности проявления деструкторов в профиле человека. HDS оценивает личность по 11 шкалам, которые соответствуют дисфункциональным личностным характеристикам.

При формировании шкал для опросника HDS Хоган взял за основу 10 расстройств личности из Диагностического и статистического руководства по психическим расстройствам, 4-е издание, пересмотренное, а также добавил шкалу для описания пассивно-аггрессивной личности, которая включена в DSM [171].

При этом, Хоган подчеркивает, что HDS не предназначен для измерения расстройств личности, а оценивает саморазрушительные или деструктивные проявления нормальной личности [171].

Также Хоган обращает внимание, что шкалы опросника HDS представляют собой своего рода спектр дисфункциональных наклонностей, возникающих в межличностных отношениях. И в зависимости от того, на какой части спектра находятся результаты конкретного человека, возможны различные поведенческие проявления: от эффективных навыков и уверенности в межличностном взаимодействии до некомпетентности и поведения, связанного с расстройством личности. Так более высокие баллы указывают на большую вероятность возникновения дисфункционального поведения в любом межличностном контексте. Хоган отмечает, что большинство людей находятся в середине этого распределения, и любой отдельный человек может иметь высокие или низкие баллы по любой из шкал [171]

### 7.13.3 Описание research gap

На основе обзора публикаций и исследований о гендерной специфике лидерства, можно сделать несколько выводов:

Исследования различий между мужчинами и женщинами - реальными лидерами и руководителями недостаточны и фрагментарны. По-видимому, получаемые данные зачастую оказываются противоречивы,

поэтому метаанализ эмпирических работ не показывает значимых отличий. Есть ряд характеристик, по которым гендерные различия у мужчин и женщин- лидеров установлены как эмпирическими исследованиями, так и подтверждаются метаанализом. В частности - гендерная идентичность, мотивация к достижениям, эффективность лидерства в однородных и смешанных группах, конкурентность/кооперативность, отказ от лидерства, самооценка.

Тема гендерной специфики деструкторов у руководителей изучена мало, этот сюжет не выделяется в тематическом репертуаре исследований, посвященных изучению гендерной психологии лидерства в целом и женского лидерства - в частности. Это связано с объективной труднодоступностью этой целевой категории и невозможностью построить сколь-нибудь масштабное (для достижения статистической значимости) и методологически обоснованное исследование. Одновременно с этим психоаналитическая традиция в изучении гендерной специфики лидерства и лидеров есть, но в контексте дисциплинарных исследований она представлена мало, теряется на фоне массы психологических работ.

Кроме того, источник затруднений мы видим также и в том, что большинство исследований используют методики, рассчитанные на когнитивную работу, или же выстраивают модели экспериментов, рассчитанных на проявление поведенческих реакций. Психоаналитическая традиция в изучении гендерной специфики лидерства и лидеров есть, но в контексте дисциплинарных исследований она представлена мало, теряется на фоне массы психологических работ.

Учитывая сказанное, мы считаем уникальной возможность изучить гендерную специфику в проявлении деструкторов у руководителей российских компаний на реальных эмпирических данных, собранных целевым образом у интересующей нас группы с помощью методики, хорошо фундированной как с точки зрения теоретической традиции, так и с точки зрения методологии личностной оценки

#### **7.13.4 Методы сбора и обработки данных**

Источником данных стала база данных результатов оценки руководителей российских компаний, собранных по методике Хогана. Для целей исследования важно, что это именно состоявшиеся руководители, занимающие достаточно высокие посты в структуре компаний, представители топ-менеджмента, которые прошли опросник HDS за период с 2019 по 2021 год. Исходный массив составлял 1743 респондента. Однако в 47% случаев наблюдались пропуски ответов респондентов по ряду ключевых характеристик исследования, включая как вопросы измерения лидерских качеств, так и социально-демографические характеристики. Всего после исключения невалидных или нерелевантных данных база содержала сведения о 922 людях из 37 компаний 9 секторов экономики.

Методическая рамка исследования базируется на нескольких принципах. Во-первых, это акцент на предварительном анализе объекта исследования на основе ряда социально-экономических переменных и переменных методики Хогана. Структура данных сформирована таким образом, что значительная часть интересующих нас переменных имеют в среднем только 2 возможные опции (мы говорим о бинарных дихотомических шкалах, о которых речь пойдет далее). Это накладывает определенные ограничения в свободе выбора методологии статистического анализа. Поскольку нас интересует идентификация и описание возможной статистической связи между компонентами деструкторов (измеряемых по методике Хогана) и рядом социально-экономических характеристик, выбор был сделан в пользу бинарной логистической регрессии, наиболее оптимальной как с точки зрения требований, так и возможностей по

верификации эффектов. Для построения серии логистических моделей предполагается выполнения ряда принципиальных шагов:

1. Разведочный анализ (EDA). На данном этапе проводится анализ используемых в модели логистической регрессии переменных, в первую очередь на предмет вероятных отклонений значений деструктивных шкал от общей тенденции. Дополнительно рассматриваемые шкалы сравниваются признаку и отрасли деятельности руководителей.
2. Формирование уравнения логистической регрессии. В первую очередь проводится оценка распределения зависимой переменной (пол респондента, отрасль деятельности) и независимых переменных (деструкторов по методике Хогана). Далее эти переменные преобразуются в бинарные (дихотомические) по следующей схеме:
  - Для параметра «пол» зависимая переменная получила значения: «0» – мужчины, «1» – женщины.
  - Каждая из 11 шкал деструкторов были перекодированы по схеме: от 0 до 70 – «0», а 70 и более – «1», что соответствует уровням проявления деструкторов у респондента. Значения выше 70 попадают в область умеренно высокого риска, значения выше 90 – в область высокого риска. Шкала считается деструктором в профиле конкретного респондента, если значение по ней составило 70 и более процентиляй.
3. Оценка статистической значимости гендера, отраслевой принадлежности руководителя и диагностируемых у него деструкторов на основе продвинутых методов логистического регрессионного моделирования. Сюда входит построение простых бинарных логистических регрессий, а также аналогичных решений с применением алгоритмов сэмплинга с использованием Марковских цепей Монте-Карло (МСМС-подход). Анализ статистической значимости подразумевает интерпретацию специализированных коэффициентов — информационный критерий Акаике (AIC), Байсовский информационный критерий (BIC), в том числе оценку статистической значимости. Далее проводится поиск и формирование системы аргументации на основе полученных результатов.

#### **7.13.5 Обсуждение**

Происходящие в стране и в мире события усиливают актуальность и значимость как нашей работы, так и любых исследований в этой области. Кризис, стресс, ситуация неопределенности провоцируют актуализацию защитных механизмов, которые проявляются как деструкторы. Первой на линии удара оказывается категория руководителей, лидеров, вынужденных действовать в этих условиях, принимать решения, поддерживать и мобилизовывать своих последователей. Деструкторы руководителей имеют потенциальную опасность для благополучия команды и эффективности ее действий, предопределяют форму регресса группы, когда она оказывается неспособной адекватно реагировать на ситуацию, анализировать ее, принимать верные решения и реализовывать их. Наше исследование позволяет углубиться в понимание природы деструкторов и их гендерной специфики.

### **7.13.6 Результаты исследования**

Результаты исследования были использованы для подготовки и защиты магистерской диссертации, легли в основу научной статьи, находящейся на завершающей стадии подготовки. Кроме того, материалы работы используются в программах компании Kontakt Intersearch.

Частотный анализ деструкторов показал:

1. В кластере «Отдаление от людей» по Хорни лидирует деструктор Скептичный.
2. В кластере «Движение против людей» по Хорни очень близкими оказываются значения сразу двух деструкторов: Самоуверенный и Увлекающийся.
3. В кластере «Движение навстречу людям» по Хорни лидирует деструктор «Исполненный сознания долга», и он же оказывается безусловным лидером среди всех остальных деструкторов.
4. Самым частотным оказывается кластер «Движение навстречу людям» по Хорни: деструкторы из этого кластера встречаются в среднем в два раза чаще, чем деструкторы по шкалам из других кластеров.

Изучение результатов регрессионного анализа позволяет зафиксировать несколько моментов:

1. Шкалы деструкторов Скептичный, Осторожный, Сам в себе, Самоуверенный, С богатым воображением и Исполненный сознания долга оказываются наиболее важными и дифференциирующими для гендерных групп.
2. Мужчинам более всего свойственны деструкторы Скептичный и С богатым воображением.
3. Для женщин характерны три деструктора: Осторожный, Сам в себе, Самоуверенный и Исполненный сознания долга.

### **7.13.7 Заключение**

Использование бинарной логистической регрессии позволило подтвердить **основную гипотезу исследования**: гендерная специфика в проявлении деструкторов у руководителей российских компаний существует. **Ключевой результат работы** в том, что мы выделили шкалы, по которым наблюдалась устойчивая связь с гендером. Ими стали: «Скептичный - Параноидный» и «С богатым воображением - Шизотипический» у мужчин; и «Осторожный - Уклоняющийся», «Сам в себе - Шизоидный», «Самоуверенный - Нарциссический», «Исполненный сознания долга - Зависимый» - у женщин.

Научная новизна заключается в выборе психоаналитической традиции в качестве модели объяснения и теоретической проработки темы. Мы продемонстрировали, что анализ эмпирических данных позволяет увидеть гендерную специфику деструкторов, выявленных у руководителей российских компаний. Это, в свою очередь, позволяет нам говорить о подтвержденной исходной гипотезе нашего исследования. Одновременно с этим опора на целый ряд классических исследований и теорий позволяет увидеть в эмпирических результатах психоаналитические основы деструкторов, проявляющихся у мужчин и женщин - руководителей.

Почему мы считаем психоаналитический подход полезным и эвристичным? Если говорить в целом, эта традиция дает возможность прикоснуться к первоосновам деструктивных моделей лидерства. Это особенно важно на фоне большого числа публикаций как о «токсичном лидерстве», так

и о гендерной дискриминации («стеклянный потолок» для женщин, претендующих на руководящие позиции), стилистика и характер которых говорит о выраженной «моральной панике». По большому счету, приоритет в таких работах отдается социальному контексту, который рассматривается в качестве ключевого фактора как деструктивного лидерства, так и связанных с ним патологий организаций.

Психоаналитический взгляд на проблему нам представляется более аккуратным и точным в своей базовой аналитической стратегии. Во-первых, он демонстрирует связь лидерских качеств (в том числе и деструкторов) с личностными характеристиками, сформировавшимися у человека в детстве под влиянием особенностей его взаимодействия с матерью и отцом. Область бессознательного становится, таким образом, еще одним объясняющим фактором - наряду с социо-культурным контекстом.

Во-вторых, он естественным образом вводит идею гендерной специфики лидерства, поскольку опирается на описанные различия в процессе формирования личностных характеристик у мальчиков и у девочек. На наш взгляд такое более глубокое и гендерно дифференцированное понимание феномена позволяет избежать примитивизации как наших аналитических стратегий, так и предложений, что следует делать, чтобы найти «быстрое и эффективное» решение как поддержать развитие женского лидерства. Например, идея «позитивной дискриминации» женщин — руководителей.

В-третьих, он позволяет увидеть природу деструкторов сразу же в их гендерной несопоставимости, асимметричности для мужчин и женщин. Генезис «мужских» деструкторов скорее предопределен ролью матери и приводит к параноидальному или шизотипическому типу их проявления в лидерстве. На формирование женских деструкторов сильное влияние оказывает роль отца, если она была «токсичной». Это чревато переносом модели отношений отца и дочери на отношения с авторитетными и властными «значимыми мужчинами».

В-четвертых, психоаналитический подход предлагает эвристичные модели объяснения эмпирически наблюдаемым фактам (наблюдению, экспериментальным и статистическим данным, материалам консультаций и коучинговых сессий), предлагает рабочий понятийный аппарат: «защитные механизмы», «системы мотивационных потребностей», «главная тема конфликтных отношений», «типы организации личности» и др.

В-пятых, психоаналитическая традиция укрепляет наше понимание, почему деструкторы проявляются именно в стрессовой ситуации, объясняет механизмы такого регресса относительно «нормальной», рабочей ситуации как на индивидуальном, так и групповом уровне.

## **8 Профессиональные роли журналистов: об исследовательском проекте**

### **8.1 Введение**

Содержание различных медиа-платформ и новостей меняется в разных культурах. Целью этой работы является лучшее понимание факторов, объясняющих различные модели журналистики во всем мире, а также разрыв между нормами, профессиональными идеалами и позициями, представленными в новостных публикациях.

Анализ основан на результатах второй волны проекта «Реализация журналистских ролей» (Journalistic Role Performance, JRP) ([www.journalisticperformance.org](http://www.journalisticperformance.org)) - международного исследования,

основанного на данных 365 средств массовой информации в 37 странах. Руководитель исследования – профессор Клаудия Мелладо (Университет Вальпараисо, Чили). С российской стороны руководителем национальной исследовательской команды выступает аналитик МЛ ПСА С.Г. Давыдов. Основной задачей второй волны исследования, которая проводилась в 2020 – 2021 гг., стал систематический анализ того, как различные профессиональные роли анализируются в новостях. Анализ журналистских ролей в печатных СМИ и проектах национальных газетных новостей был проведен в рамках первой волны исследования, которая проводилась в период с 2013 по 2017.

Исследование базируется на комплексном подходе. Основываясь на стандартизированной операционализации ролей «сторожевого пса», интервенциониста, лояльного посредника, гражданской, информационно-развлекательной и обслуживающей ролей в журналистике, сначала было проведено измерение уровня представленности той или иной журналистской роли в новостных текстах посредством контент-анализа. Для выявления связи между оценочным и перформативным уровнями журналистской культуры далее был проведен опрос о концепции роли, ее восприятии и воплощении среди журналистов, работающих в новостных медиакомпаниях, включенных в нашу выборку. Это позволило сравнить оценки журналистов со средними показателями работы их новостных СМИ по странам. Каждая национальная команда собирала информацию на организационном/институциональном уровне своих средств массовой информации и на уровне стран. В исследовании используется методическая схема, основанная на сравнительном изучении демократических, переходных и недемократических стран. Журналистская практика встроена в рутину и осуществляется в рамках социальной системы, которая служит основой для создания медиаконтента. Таким образом, возникает возможность глубоко изучить профессиональные роли, как различные социальные и культурные контексты объясняют различия в воплощении журналистских ролей для разных медиаплатформ и тем медийной повестки дня.

Полученные данные были проанализированы посредством методологии контент-анализа. Затем к полученным результатам была применена методология сетевого анализа для построения сети со-встречаемости журналистских ролей.

Полученные результаты показывают, что осуществляемые в практической деятельности журналистов роли в значительной степени отличаются от «идеальных типов», сформулированных и описанных в теоретических работах по медиа и журналистике. Результаты сетевого анализа показывают связи как между ролями, так и между отдельными индикаторами конкретных ролей, что позволяет делать выводы об их целостности и пересечению (наслоению) друг с другом.

## 8.2 Методы сбора и обработки данных

**Сбор данных.** Стремясь получить гетерогенную выборку, авторы исследования выбрали страны, которые представляют различные политические режимы, географические регионы и классификации медиасистем. В исследование вошли государства Северной Америки, Латинской Америки, Западной Европы, Восточной Европы, Азии, Африки, Ближнего Востока и Океании. Следуя моделям западных медиасистем Д. Халлина и П. Манчини, в выборку вошли страны, которые представляют либеральную, демократическую корпоративистскую и поляризованную плюралистическую модели. Также были использованы индексы демократии и отчеты о свободе прессы для классификации демократических стран с переходной экономикой и недемократических стран из разных частей мира.

Данные были собраны соответствующими национальными исследовательскими коллективами в 37 странах – Аргентине, Австралии, Австрии, Бельгии, Бразилии, Канаде, Чили, Колумбии, Кубе, Эквадоре, Египте, Великобритании, Эстонии, Эфиопии, Франции, Германии, Венгрии, Ирландии, Израиле, Италии, Японии, Кувейте, Ливане, Мексике, Парагвае, Польше, Катаре, России, Руанде, Сербии, Испании, Швейцарии, Тайване, Объединенных Арабских Эмиратах, США и Венесуэле.

Полевые работы по контент-анализу, онлайн-опросу и сбору данных на организационном и социальном уровне проводились в период с 2020 по 2021 год. Все данные контент-анализа, опроса и измерения структурного контекста были собраны независимо. Это позволило организовать работу на разных уровнях анализа: новости, журналисты, средства массовой информации и страны. **Обработка данных контент-анализа.** В выборку контент-анализа вошли новости, опубликованные в крупнейших газетах, на сайтах, в новостных радио- и телепрограммах стран-участниц исследования.

Хотя для отбора конкретных текстов внутри каждой страны использовались разные процедуры в зависимости от типа анализируемой медиаплатформы, все они были связаны общими техническими аспектами, чтобы гарантировать возможность использования глобальной выборки для проведения сравнений. Эти аспекты связаны с общим периодом времени, для которого были взяты тексты для анализа, одинаковыми днями анализа и с одними и теми же единицами анализа для всех стран. Была использована следующая модель построения выборки исследования. Исследователи в каждой из 37 стран-участниц выбрали от двух до четырех новостных СМИ на каждую платформу. Критериями выбора средств массовой информации были размер аудитории, охват и уровень влияния на формирование национальной повестки дня. Исследователям было предложено отобрать наиболее популярные СМИ (наиболее читаемые, просматриваемые или прослушиваемые) в своем классе на основе рейтингов или аналогичных параметров, и предпочтение отдавалось медиа национального масштаба, хотя в исследование были включены также региональные и местные медиа площадки.

Учитывая, что структура и формат медиасистем в разных странах во многом различаются, включая размер, ориентацию аудитории, собственность в сфере медиа, политические взгляды и наличие более чем одного языка на территории, исследователей попросили убедиться, что выбранные средства массовой информации представляют разнообразие медиа каждой страны в максимально возможной степени. Членам исследовательских команд пришлось принять во внимание тот факт, что количество включенных средств массовой информации может варьироваться от страны к стране и что большая гетерогенность в системе средств массовой информации приведет к тому, что исследователи будут включать в выборку больше средств массовой информации, и наоборот.

Чтобы контролировать потенциальную перепредставленность и/или недопредставленность конкретных типов СМИ в выборке, возникающую из-за того, что некоторые средства массовой информации включают в выборку больше материалов, чем другие, было принято решение взвесить данные по средствам массовой информации для каждой страны. Это гарантировало, что каждый тип СМИ – телевидение, радио, онлайн-новости и газеты – в каждой стране будут иметь одинаковый вес в обобщенных результатах. Количество проанализированных медиа и новостей в каждой стране представлено в Таблице ??.

Количество новостных текстов и СМИ в выборке исследования по странам

<b>СТРАНА</b>	<b>НОВОСТНЫЕ ТЕКСТЫ</b>	<b>КОЛИЧЕСТВО СМИ</b>
Аргентина	5368	10
Австралия	1965	8
Австрия	4821	10
Бельгия	2411	7
Бразилия	3679	9
Канада	3727	12
Чили	7512	11
Колумбия	5138	8
Эквадор	2892	8
Египет	3484	11
Великобритания	4185	15
Эстония	2409	11
Эфиопия	1400	10
Франция	4661	9
Германия	4777	6
Венгрия	3358	13
Ирландия	2421	8
Израиль	2448	8
Италия	4494	11
Япония	3757	9
Кувейт	1868	9
Ливан	3665	14
Мексика	7905	12
Польша	6230	14
Катар	1559	6
Россия	6955	10
Руанда	2644	7
Сербия	6067	11
Южная Корея	3959	8
Испания	6089	12
Швейцария	3555	10
Тайвань	6790	10
ОАЭ	2726	9
США	3992	11
Венесуэла	2443	12
Куба	2834	8
Парагвай	4286	8

---

Количество новостных текстов и СМИ в выборке исследования по странам

---

ВСЕГО

148,474

365

---

С помощью метода «сконструированной недели» была сформирована стратифицированно-систематическая выборка размерностью две недели для каждого средства массовой информации в каждой стране со 2 января по 31 декабря 2020 года. Во всех странах, включенных в исследование, были проанализированы одни и те же дни. Поскольку ежедневные и ежемесячные колебания публикационной активности являются важными факторами, которые следует учитывать при проведении анализа контента новостей, мы разделили год на два периода (полугодия) по шесть месяцев: с января по июнь и с июля по декабрь. Затем внутри этих периодов по определенному алгоритму были отобраны дни. Алгоритм выбора дней подробно представлен на сайте проекта <https://www.journalisticperformance.org/>. В результате в каждом из двух периодов (полугодий) была сконструирована неделя из неповторяющихся дней для проведения измерений, и итоговая выборочная совокупность в исследовании составила 14 дней одного года. Использование такого алгоритма позволило достигнуть презентации по месяцам (в каждом месяце выбран хотя бы один день для измерения) и медиа-источникам (в каждый из дней хотя бы одна из программ/выпусков/изданий) для их более равномерного представления.

Исследовательская команда в каждой из стран-участниц проекта использовала такие критерии для отбора единиц наблюдения в отобранных СМИ: - для новостей на выбранных телеканалах: выпуски с наибольшим количеством просмотров; - для новостей на выбранных радиоканалах: программы с наибольшим количеством прослушиваний; - для отобранных печатных (газет): полный газетный выпуск; - для отобранных онлайн средств массовой информации (сайтов): полная заглавная страница и все идущие с нее ссылки.

Поскольку онлайн средства массовой информации имеют возможность публиковать информацию не только в заранее определенное время (как другие СМИ), а по мере ее поступления, замеры на сайтах проводились в две определенные исследователями точки времени: в 11:00 и 23:00. Этот 12-часовой перерыв между двумя снимками с высокой долей вероятности обеспечивает наибольшую вариативность контента. Домашние страницы и соответствующие ссылки открывались и сохранялись в режиме реального времени.

Единицей анализа была новость. Новость определялась как совокупность связанных элементов из слов, аудио-записей и визуального материала, аффилированных с определенным происшествием/вопросом/персоной. Для каждой из отобранных единиц анализа собирались все новости по широкому кругу тем (правительство и законодательные органы, выборы и предвыборные кампании, протестная активность, экономика и бизнес, полиция и преступность, судебная деятельность, оборона и безопасность, образование, здоровье, проблемы социальной сферы, окружающая среда, энергетика, транспорт, жилье, происшествия и несчастные случаи, религия и церковная деятельность, занятость и трудовая деятельность, медиа-среда, спорт, наука и технологии, образ жизни, культурный досуг и развлечения, жизнь знаменитых людей). Затем тематические позиции были перекодированы в широкие темы.

Об одном событии, проблеме или явлении средство массовой информации может сообщать более чем в одной новости в один и тот же день. Если одно и то же событие, явление или проблема освещались

более чем в одном материале, они считались отдельными историями и кодировались отдельно.

Редакционные статьи, колонки мнений, прогнозы погоды, гороскопы, обзоры фильмов (или других культурных событий), головоломки, страницы в социальных сетях и аналогичный контент на радио и телевидении не были включены в исследование. Авторы проекта исключили приложения/журналы/специальные программы и заголовки на первых полосах газет и в начале теле- и радиовыпусков. В целях сравнения данных опроса о ролевых концепциях журналистов с содержанием их новостных организаций также был исключен контент, который не был создан сотрудниками соответствующих редакций — например, статьи новостных агентств или статьи не журналистов, размещенные на новостных сайтах. В случае онлайн-СМИ кодировались только те новости, которые появлялись на расширенной домашней странице. Также были закодированы элементы, включающие встроенные видео или аудиоклипы.

В целом, выборка состоит из 148 474 новостей, опубликованных 365 СМИ (102 газет, 96 телевизионных выпусков новостей, 74 новостных радиопрограмм и 93 новостных веб-сайтов).

**Построение сети со-встречаемости журналистских ролей.** На основе имеющихся данных может быть создана сеть, основанная на связях между публикациями и журналистскими ролями, на основе которой может быть построена сеть связей между журналистскими ролями на основе их совместного присутствия в рассматриваемых и кодируемых в рамках исследования публикациях.

Сетевые данные были представлены в формате несимметричной матрицы, где по строкам были указаны все наблюдения, а по столбцам – индикаторы, соответствующие изучаемым показателям (журналистским ролям). Фрагмент такой матрицы на примере информации о тематиках, раскрывающихся в статьях, приведен на Рис. ???. С помощью кода, написанного на языке программирования Python, матрицы была конвертированы в сетевые файлы в виде двумодальных сетей с расширением .net. Пример конвертации данных в формат 2-модальной сети приведен на Рис. ???. На имеющихся данных были построена 2-модальная сеть (сети, состоящие из двух наборов узлов, где связи могут существовать только между узлами из разных наборов) публикаций (Publications) и ролей (Roles) журналистов PR.

Создание 1-модальных сетей из 2-модальных сетей производилось через сетевое перемножение. Эта процедура подразумевает, что двумодальная сеть, представленная в виде матрицы, перемножается на свою транспонированную версию, и один из наборов узлов становится посредником для формирования связей между узлами другого набора. В частности, в данном случае была построена сеть со-встречаемости журналистских ролей RR, где вес связей указывал на количество раз, когда эти сущности встречались вместе в одном закодированном материале (статье). Анализ полученных сетевых данных проводился в программе для анализа и визуализации больших сетей Pajek.

Номер узла	148475	148476	148477	148478	148479	148480	148481	148482	148483	148484	148485	148486	148487	148488	148489	148490	148491	148492	148493	148494	148495	148496	148497	148498
Номер узла/ID наблюдения и категория	GOVERNMENT	CAMPAIGNS	POLICE	COURT	DEFENSE	ECONOMY	EDUCATION	ENVIRONMENT	ENERGY	TRANSPORTATION	HOUSING	ACCIDENTS	HEALTH	RELIGION	LABOR	PROTEST	SOCIALISSUE	MEDIA	SPORT	SCIENCE	LIFESTYLE	CULTURE	ENTERTAINMENT	OTHER
1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
3	3	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	4	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
5	5	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	7	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0
9	9	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
11	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
12	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
13	13	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
14	14	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	15	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	16	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0
17	17	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
19	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
21	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
22	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
23	23	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
24	24	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
25	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
26	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
27	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
28	28	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
29	29	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0
30	30	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0

Узлы в 1-mode (до 148,474)		Узлы в 2-mode (с 148,475)		Связи между узлами	
*Vertices	148498 148474	148473	148473	*Arcs	
1	1	148474	148474	1	148477
2	2	148475	"GOVERNMENT_LEGISLATURES"	2	148486
3	3	148476	"CAMPAIGNS_ELECTIONS"	3	148488
4	4	148477	"POLICE_CRIME"	4	148485
5	5	148478	"COURT"	5	148482
6	6	148479	"DEFENSE_SECURITY"	6	148491
7	7	148480	"ECONOMY"	7	148476
8	8	148481	"EDUCATION"	7	148488
9	9	148482	"ENVIRONMENT"	7	148475
10	10	148483	"ENERGY"	8	148489
11	11	148484	"TRANSPORTATION"	8	148491
12	12	148485	"HOUSING"	9	148477
13	13	148486	"ACCIDENTS"	10	148498
14	14	148487	"HEALTH"	10	148495
15	15	148488	"RELIGION"	11	148496
16	16	148489	"LABOR_EMPLOYEMENT"	11	148497
17	17	148490	"PROTEST"	12	148493
18	18	148491	"SOCIAL_ISSUES"	13	148477
19	19	148492	"MEDIA"	13	148491
20	20	148493	"SPORT"	13	148490
21	21	148494	"SCIENCE_TECHNOLOGY"	14	148480
22	22	148495	"LIFESTYLE"	14	148476
23	23	148496	"CULTURE"	15	148475
24	24	148497	"ENTERTAINMENT_CELEBRITY"	15	148483
25	25	148498	"OTHER"	15	148480

## Результаты исследования Базовые характеристики полученных сетей. Двумодальная сеть публикаций и ролей PR состоит из 6 ролей и 67 индикаторов ролей. Количество ролевых индикаторов по каждой роли приведено в Табл. ?.?. Роль «сторожевого пса» содержит значительно большее количество индикаторов – 30 – по сравнению с другими ролями, куда согласно дизайну исследования включено от 5 до 10 индикаторов. Визуализация ролей и индикаторов представлена на Рис. 26.

#### Глобальные роли и ролевые индикаторы

N	Global role	N of indicators	Freq%
1	WATCHDOG	30	45
2	LOYAL-FACILITATOR	10	15
3	CIVIC	10	15
4	SERVICE	7	10
5	INTERVENTIONIST	5	7.5
6	INFOTAINMENT	5	7.5
Всего		67	100

Некоторые ролевые индикаторы, которые относятся к различным журналистским ролям, чаще других встречаются в публикациях. Таблица ?? показывает, что наибольшее число упоминаний имеют

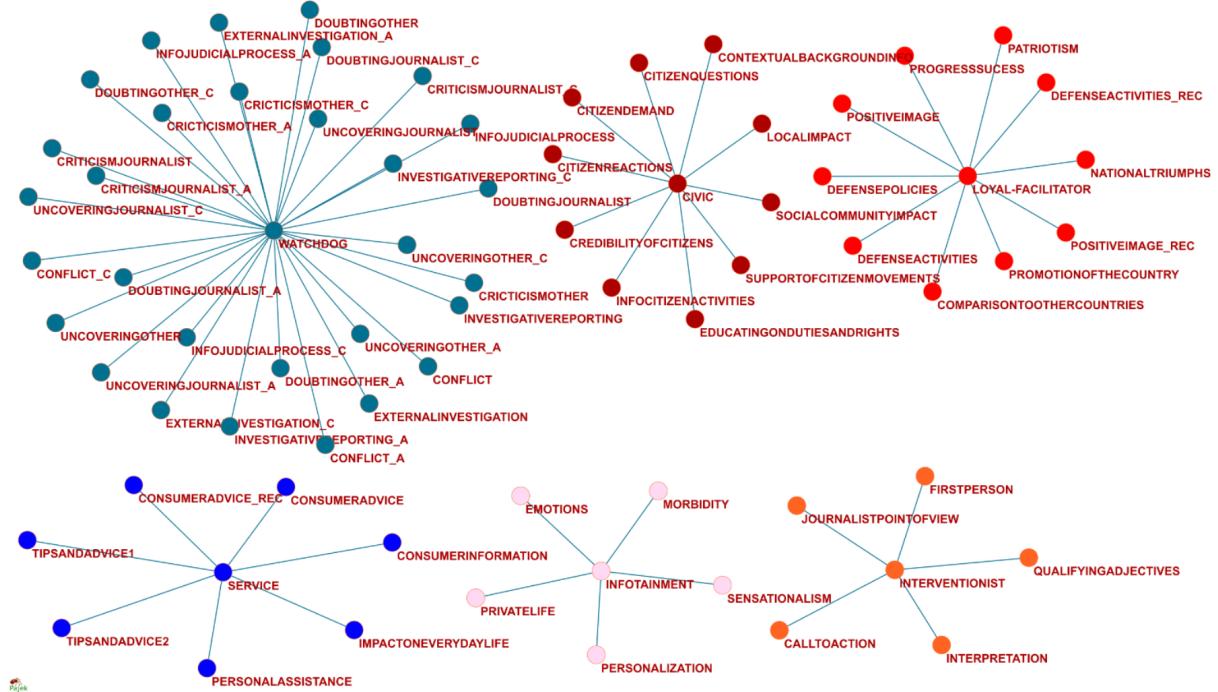


Figure 26: Визуализация сети публикаций и ролей PR

такие ролевые индикаторы как QUALIFYINGADJECTIVES (квалифицирующие прилагательные) и INTERPRETATION (интерпретация), которые встречаются заметно чаще остальных индикаторов (50,297 и 42,687 раз соответственно). Следующие за ними пять индикаторов встречаются в 20+ тыс. публикаций.

---

Ролевые индикаторы с  
наибольшим количеством  
упоминаний в публикациях

---

#	Value	Id
1	50297	QUALIFYINGADJECTIVES
2	42687	INTERPRETATION
3	26850	PERSONALIZATION
4	25868	CRITICISMOTHER
5	24302	JOURNALISTPOINTOFVIEW
6	23330	CRITICISMOTHER_A
7	21853	EMOTIONS
8	19274	SENSATIONALISM
9	17113	INFOJUDICIALPROCESS
10	16387	IMPACTONEVERYDAYLIFE
11	15203	LOCALIMPACT
12	13347	CONSUMERINFORMATION
13	13248	CONTEXTUALBACKGROUNDINFO
14	12724	SOCIALCOMMUNITYIMPACT
15	12623	CITIZENREACTIONS

---

Ролевые индикаторы с  
наибольшим количеством  
упоминаний в публикациях

---

16	11939	DOUBTINGOTHER
17	11802	PRIVATELIFE
18	11331	POSITIVEIMAGE_REC
19	11331	POSITIVEIMAGE
20	11075	INFOJUDICIALPROCESS_A
21	10675	DOUBTINGOTHER_A

---

Одномодальная сеть совстречаемости журналистских ролей RR состоит из 67 узлов и 2,195 связей между ними; при этом сила связей между узлами варьируется от 1 до 23,458. Особенность этой сети состоит в том, что она является плотно связанной – практически все узлы связаны друг с другом, и плотность составляет 0.98. Стратегия анализа сетей такого рода заключается в обрезке наименее сильных связей ниже определенного порогового значения с помощью алгоритма line cut для того, чтобы в сети между узлами остались наиболее сильные связи. На Рис. 27 представлена визуализация полученной сети в двух вариациях: с пороговым значением веса связей в 1,000 и 1,500 (количество публикаций, в которых два индикатора ролей встречались совместно).

Для того, чтобы видеть связи внутри и между группами ролевых индикаторов, соответствующие им узлы в сети сгруппированы по ролям. Можно видеть многочисленные связи между группами (по цепочке) синих, красных, оранжевых и розовых узлов, то есть ролями WATCHDOG («Сторожевой пес»), CIVIC, INTERVENTIONALIST и INFOTAINMET, а также синих и оранжевых - WATCHDOG и INTERVENTIONALIST. Индикаторы внутри двух оставшихся ролей – LOYAL-FASILITATOR и SERVICE – тоже связаны с индикаторами других ролей, однако в менее сильной степени; обращает на себя внимание отсутствие связей между индикаторами из группы CIVIC и LOYAL-FASILITATOR.

Эти выводы становятся еще более видными после укрупнения этой сети до уровня ролей и связей внутри них (представленных в виде петель) и между ними (Рис. ??). Можно обратить внимание, что сеть представляет собой почти полную клику, за исключением отсутствующих связей между ролями CIVIC и LOYAL-FASILITATOR.

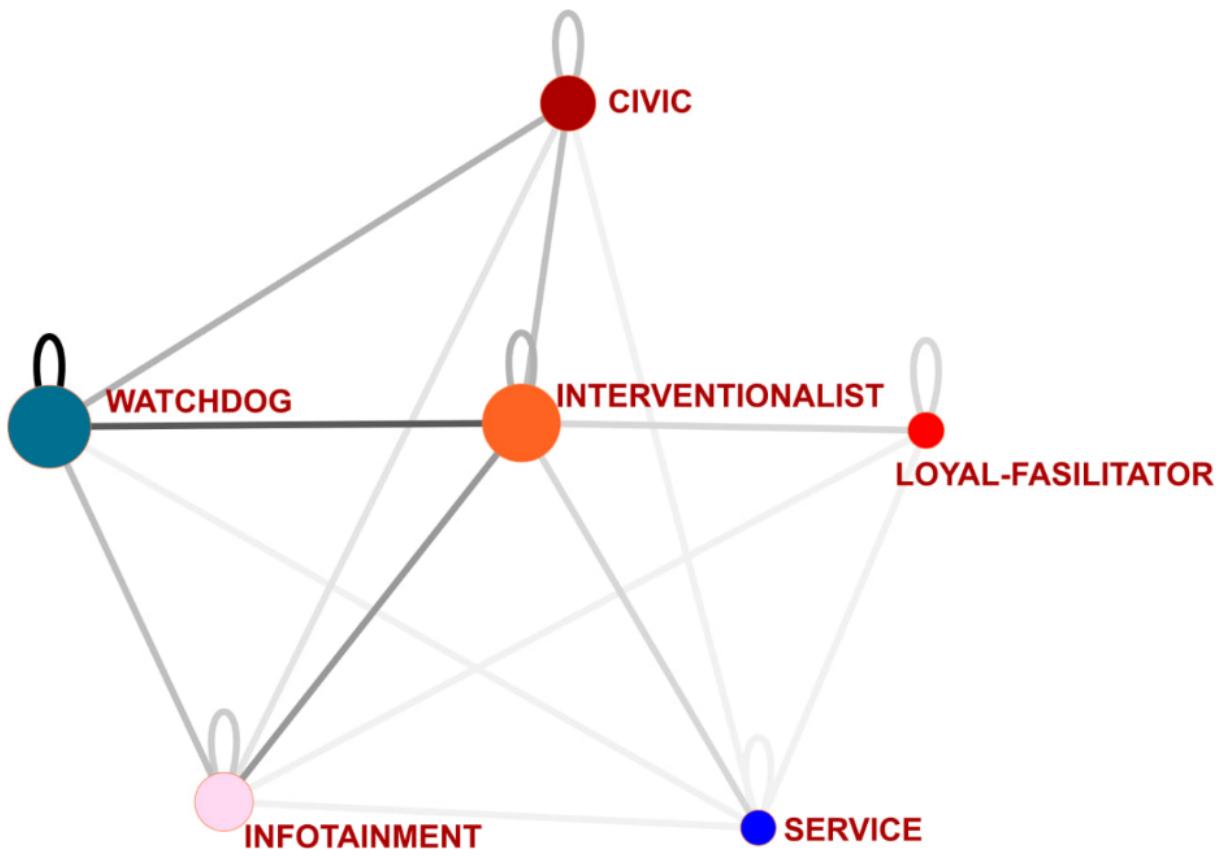
**Сетевая кластеризация - блокмоделинг.** Репрезентация сети в виде матрицы с удачным порядком расположения узлов уже может демонстрировать разделение сети на группы и показать пересечения между узлами, относящимися к разным категориям. Матричная репрезентация сети связей между индикаторами ролей по со-встречаемости RR представлена на Рис. 28. Матрица симметрична по столбцам и строкам, где представлены индикаторы журналистских ролей по 6 ролям (отмечены разным цветом). Темно-синие линии делят матрицу на группы (кластеры). Кластеры по диагонали включают индикаторы, относящиеся к одной роли, и показывают связи между ними. Кластеры вне диагонали показывают связи между индикаторами разных групп. Можно видеть наличие сильных связей внутри оранжевой и коричневой группы индикаторов, между ними также наблюдаются связи. Внутри большой группы индикаторов, относящихся к роли «Сторожевого пса», индикаторы связаны друг с другом неравномерно – тогда как у некоторых есть много связей с другими, часть индикаторов

занимают довольно изолированную позицию. Однако более точным для выделения групп и определения связей между ними является вариант сетевой кластеризации – блокмоделинг. В работе был применен непрямой подход к блокмоделингу, в ходе которого сначала рассчитывается мера близости / дальности между узлами на основании их структурного положения, на основе которой узлы затем относятся к тому или иному кластеру. В ходе анализа для расчета несходства элементов (dissimilarity) использовалось скорректированное Евклидово расстояние, которое учитывало различные веса связей в сети (вариант d5 в программе Pajek). На основе полученной дендрограммы (Рис. 29) было сделано предположение о наличии в сети трех кластеров, матрица с которыми показана на Рис. 8. Можно видеть, что наиболее плотно связанный кластер (справа внизу) состоит из восьми ролевых индикаторов. Второй кластер из 32 индикаторов также имеет достаточно много связей внутри себя, а также сильно связан с кластером 1. Индикаторы, вошедшие в кластер 3, не имеют значительного числа связей с индикаторами из своей группы, а также имеют выборочные связи с индикаторами из групп 1 и 2. Чтобы понять, к каким журналистским ролям относятся индикаторы, вошедшие в три кластера, для каждого лейбла был задан цвет, к которому он относится исходя из принадлежности к определенной роли (Рис. ??, справа). Полученные результаты позволяют говорить о том, что похожие друг на друга по структурным характеристикам кластеры (в смысле соотношения индикаторов друг с другом) на деле образованы индикаторами, которые относятся к разным журналистским ролям. Так, наиболее плотно связанный кластер 1, где имеется много связей между индикаторами, формируется индикаторами из ролей, обозначенных коричневым (гражданская журналистика), оранжевым (интервенционист) и синим («сторожевой пёс»).

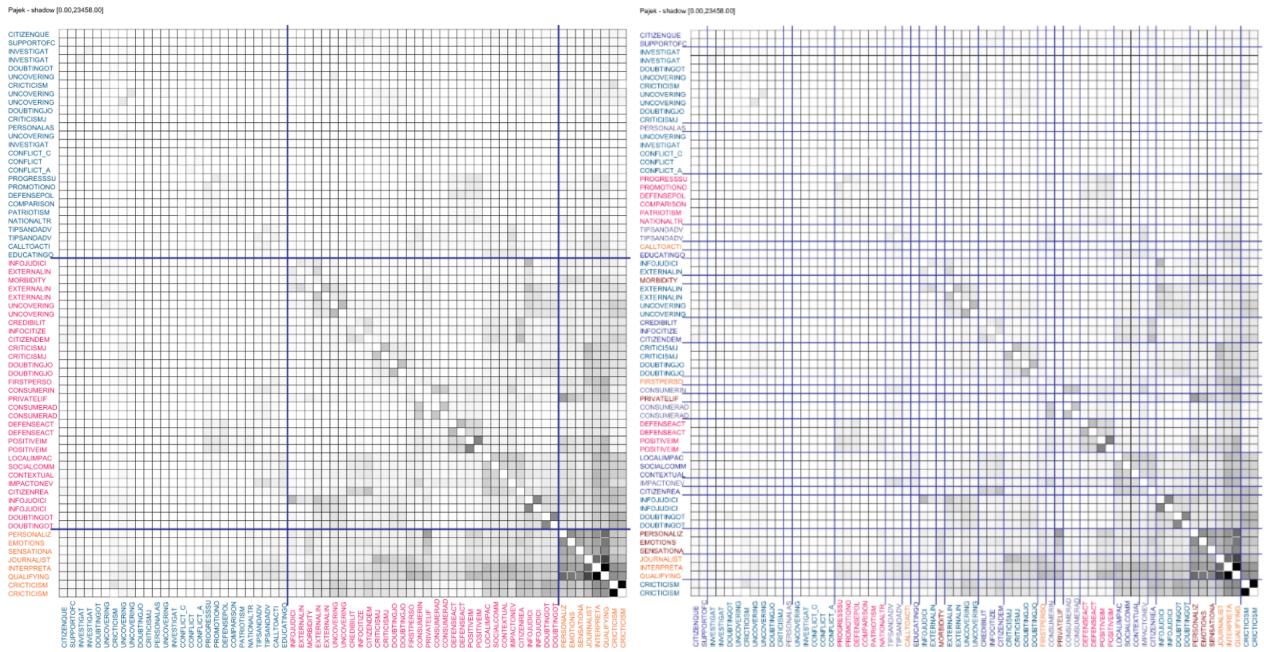
Для визуализации веса связей нормализованы по максимальному значению. Размер узла приведен в соответствие с взвешенным показателем входящей центральности (weighted degree). Цвет узлов соответствует разделению на группы по 6 ролям.



Figure 27: Визуализация сети связей между индикаторами ролей по со-встречаемости RR (пороговые значения по весу связи 1000, 1500)



Укрупненные узлы по ролям расположены на тех же местах графа, что и группы узлов на Рис. 27.



## Обсуждение и выводы Полученные результаты показывают, что осуществляемые в практической деятельности журналистов роли в значительной степени отличаются от «идеальных типов», сформулированных и описанных в теоретических работах по медиа и журналистике. Оказалось, что роли интервенциониста, «сторожевого пса» и гражданско-журналиста в значительной степени пересекаются по своим

Pajek - shadow [0.00,23458.00]

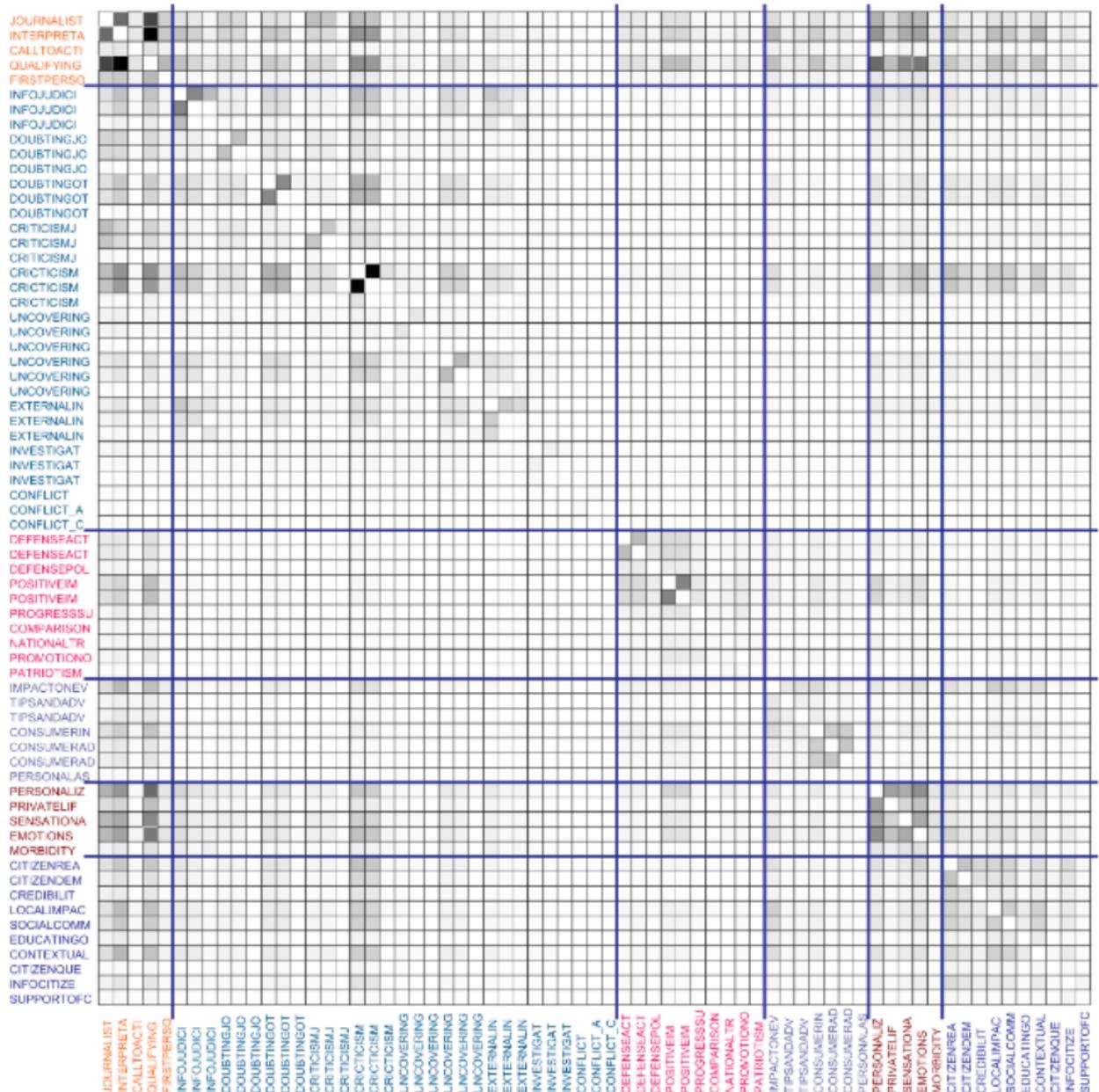


Figure 28: Визуализация сети связей между индикаторами ролей по со-встречаемости RR в виде матрицы

## Pajek - Ward [0.00,295056.88]

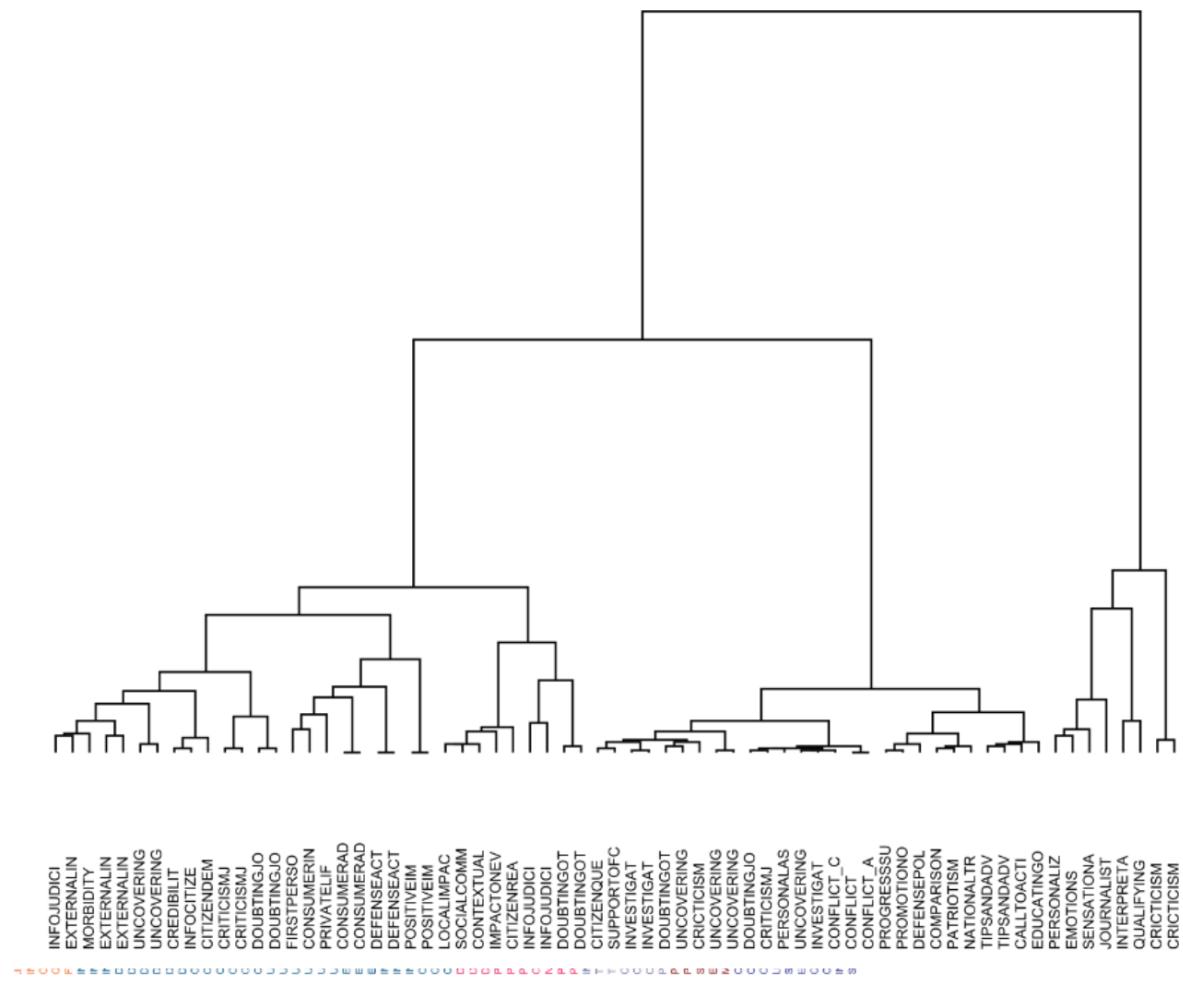


Figure 29: Дендрограмма по расчетам скорректированного Евклидова расстояние

индикаторам. В наименьшей степени пересекаются с другими ролями индикаторы лояльного фасилитатора.

Наиболее целостной, если судить по пересечению индикаторов внутри одной роли, является роль «сторожевого пса». У остальных ролей данный показатель заметно ниже, слабее всего внутренние связи выражены у сервисной роли. Самая слабая связь наблюдается между ролями гражданского журналиста и лояльного фасилитатора, что вполне ожидаемо, поскольку первый отстаивает интересы гражданских сообществ, тогда как второй в первую очередь поддерживает государственную бюрократию и находящиеся у власти элиты.

### **8.3 5. Обеспечение практического внедрения инструментов в прикладные аналитические и консалтинговые проекты**

#### **8.3.1 2. Возможности и ограничения анализа открытых судебных данных: кейс данных по административным делам по протестам на сайте Московского Городского Суда**

##### **8.3.2 Введение**

В этой части отчета описаны возможности и ограничения анализа открытых судебных данных, сформулированные по итогам анализа текстов судебных решений по делам о протестах (статья 20.2 КоАП РФ) в Москве с использованием данных с сайта Московского Городского Суда. Сперва представлены основные положения правового регулирования публикации текстов судебных решений в сети “Интернет,” далее описаны способы получения доступа к открытым судебным данным и технологии, используемые в их обработке. Обзор закрывает описание возможностей и ограничений анализа данных на текстах судебных решений, а также оценка доступности данных и описание встреченных процедурных нарушений.

Несмотря на постепенное сужение списка доступных открытых государственных данных в России (в сравнении с политикой публикации данных отдельных государственных органов в начале-середине 2010-х годов), судебные данные в России начали публиковаться публично в 2011 году и доступны на сайтах судов или в системе ГАС “Правосудие” [15]. Публикация текстов судебных решений регулируется Федеральным Законом №262 [6], а отдельные сроки прописаны в Постановлении Президиума Верховного Суда Российской Федерации: согласно им, судебные акты судов общей юрисдикции должны публиковаться в “разумные сроки”, но не позднее месяца с даты принятия судебного акта [10]

##### **8.3.3 Сбор открытых судебных данных**

Есть несколько способов получения доступа к открытым судебным данным. Во-первых, можно выбрать – использовать ли сторонний сервис, который предоставит данные по судебным решениям за определенную плату, или собрать данные самостоятельно с использованием технологии web-scraping, которая реализуется преимущественно на

языке Python. Современные методы сбора данных из открытых источников позволяют исследователям из области социальных наук с базовым уровнем навыков в области программирования собирать данные самостоятельно; растет корпус исследований текстов судебных решений и самостоятельно

собранных и пред обработанных судебных решений [459], [16]. Во-вторых, можно выбрать либо отдельный сайт суда и организовать сбор данных на нем, либо собирать данные из Государственной Автоматизированной Системы “Правосудие”, на которой публикуются данные федеральных судов. Третья развилика следующая – собирать все данные с веб-страниц или только скачивать тексты судебных решений. В рамках нашего исследования мы выбрали путь сохранения максимального количества метаданных об административном деле, поэтому пошли по пути сохранения всех данных. Итак, итоговый подход к сбору данных выглядел так: мы собирали данные с помощью веб-скрапера, написанного на языке Python для сбора данных с сайта Московского Городского Суда. Технология веб- скрапинга (web-scraping) представляет из себя автоматизированный сбор данных по заранее написанным правилам, поэтому данные собирались автоматизированно [269]. Итоговым результатом стала выгруженная база данных о 35 615 административных делах с сайте Московского городского суда (все годы, все инстанции).

Дальнейшая работа была связана с предобработкой данных. Во-первых, следует выбрать уровень судебной инстанции (первая инстанция, апелляция, кассация), установить ограничения по году если необходимо. Мы сосредоточились на судебных решениях первой инстанции с 2017 года в Москве. Далее необходимо написать часть функций, основывающихся на регулярных выражениях и логических правилах, чтобы вычленить детали административного дела из текста судебного решения: например, по присутствию в тексте таких слов как “административный арест”, “административный штраф”, “обязательные работы” можно понять, какая мера наказания была назначена обвиненному. Вычленение социально-демографических характеристик обвиняемого и подробностей дел основывается на регулярных выражениях, логических функциях, и результате распознавания именованных сущностей с помощью библиотеки Natasha для Python. Также, можно сделать вывод о части социально-демографических характеристик обвиняемого (как минимум, указывается семейный статус, наличие иждивенцев, инвалидности, адвоката, работы, официального места жительства, и др.). Итоговый набор данных будет иметь следующую структуру – 1 строка представляет из себя 1 административное дело. В нашем случае – по статье 20.2 КоАП РФ в г. Москва в период с 2017 гг. по 2023г., завершенное на момент сбора данных, по которому присутствовал текст судебного решения первой инстанции.

Полноту административных данных по статье 20.2 КоАП РФ на сайте Московского городского суда можно описать следующим образом:

- Сбор ссылок на страницы дел: 99.7-100% данных доступно;
- Сбор данных на страницах дел: 100% данных доступно;
- Отсутствуют какие-либо судебные документы за последние 5 лет в среднем – 6.95% дел;
- Зашифрован размер штрафа – в 50% случаев, где назначается штраф.

#### 8.3.4 Ограничения и возможности

Основные сложности, с которыми мы столкнулись, связаны с процедурными нарушениями при публикации текстов судебных решений, а также с тем, что многие обвиняемые не приходят или не могут прийти на судебное заседание, из-за чего данные об их образовании, месте работы и прочие характеристики не указываются. Следует учитывать, что еще в 2012 году анализ судебных

актов, опубликованных в интернете сотрудниками аппаратов судов показал, что в открытом доступе публикуется примерно половина судебных актов [472]; с годами эта тенденция скорее исправлялась в сторону публикации большего процента судебных решений, но даже опубликованные акты зачастую лишены информации, не подлежащей удалению – Михаил Поздняков отмечает, что зачастую из текст судебных актов удаляется номер акта. Мы обнаружили, что в судебных актах зачастую не указано имя судьи, не указано имя адвоката, размер штрафа и иногда зашифровано даже число суток ареста. Другой пример процедурного нарушения – публикация персональной информации об обвиняемом: встречаются постановления о назначении административного наказания с указанием полного адреса проживания потерпевшего. Все это противоречит Постановлению Президиума Верховного Суда РФ от 27.09.2017, что регулирует порядок публикации судебных актов в сети “Интернет” [10]. Однако, следует отметить и возможности, которые открываются при работе с административными данными из открытых судебных источников – во-первых, эти данные публикуются в открытом доступе и легко поддаются автоматизированной обработке с помощью Python или других языков программирования. При наличии социально-демографических характеристик обвиняемых также не представляет сложности вычленить их из текста. Однако, нужно учитывать и то что в части случаев социально-демографические данные указаны либо в усеченном виде, либо не указаны вообще.

Возможности: 1. Публикуются в открытом доступе, легко поддаются автоматизированной машинной обработке; 2. В части текстов судебных решений доступны социально-демографические характеристики обвиняемых; 3. Высокий процент доступности текстов судебных решений;

Ограничения: 1. Часть данных не опубликована ([472]); 2. Часть социально-демографических характеристик обвиняемых не указана, в других случаях характеристики не указываются из-за того что обвиняемые не участвуют в судебном разбирательстве; 3. Часть данных текстов судебных решений зашифрована.

Проведенное исследование вносит свой вклад в понимание доступности судебных решений и текстов судебных решений в России. Административное судопроизводство даже в крупных российских судах не выделено в рамки отдельных коллегий (как в случае с уголовными и гражданскими делами); нет и судей, специализирующихся преимущественно на административных делах – такими делами занимаются в основном судьи, специализирующиеся на гражданском судопроизводстве (Волков и др., 2015). Нет и стимулов, по которым внимание к техническому сопровождению дел со стороны сотрудников аппарата суда было повышенным – нет других заинтересованных сторон, как при наличии прокуратуры в случае уголовного судопроизводства или истца и ответчика в гражданском процессе. Аппараты судов работают в условиях большой нагрузки и низкой зарплаты труда (особенно относительно больших городов). Но так или иначе, социально-правовые исследования с использованием административных данных возможны, хотя и требуют учитывать ряд определенных юридических и процедурных ограничений.

## **8.4 5.3 Разработка и бизнес-применение инструментария для оценивания репутации брендов на основе семантического сетевого анализа**

### **8.4.1 Введение**

Оценка эффективности, восприятия и позиции бренда важна для различных функций коммерческих компаний: управление брендом и маркетингом, управление клиентским опытом, развитие корпоративной

культуры и многие другие.

Поскольку социальные сети стали важной площадкой для формирования имиджа бренда [211, 381] крайне важно разработать специальные показатели, которые могли бы помочь оценивать положение и восприятие брендов в онлайн-среде.

Цель данного проекта — описать и проанализировать как сетевые, так и традиционные показатели, используемые для оценки эффективности, восприятия и позиции бренда в социальных сетях. Кроме того, мы стремились описать существующие подходы, которые могут интегрировать сетевые и традиционные показатели.

#### 8.4.2 Подходы, основанные на использовании отдельных индексов

Обзор литературы показал, что исследователи крайне редко предпринимают попытки комплексного обзора показателей, связанных с анализом брендов в социальных медиа. На основе отбора избранных статей [29, 34, 130, 307, 317] мы выбрали и сгруппировали восемь ключевых показателей, которые используются для оценки эффективности, восприятия и позиции бренда (Таблица 1).

Табл. 1. Показатели оценки эффективности, восприятия и позиции бренда

КЛАССИЧЕСКИЕ		СЕТЕВЫЕ
ОБЪЕКТ АНАЛИЗА — ПОСТЫ И РЕАКЦИИ		ОБЪЕКТ АНАЛИЗА — СЕТЬ СВЯЗЕЙ МЕЖДУ СЛОВАМИ В ПОСТАХ
<b>МЕТРИКИ</b>	<b>МЕТРИКИ</b>	<b>МЕТРИКИ ПОЗИЦИИ</b>
<b>ЭФФЕКТИВНОСТИ*</b>	<b>ВОСПРИЯТИЯ</b>	
КОЛИЧЕСТВО УПОМИНАНИЙ	ТОНАЛЬНОСТЬ (SENTIMENT)	СВЯЗАННОСТЬ (CONNECTIVITY)
ОХВАТ (REACH)	ЛОЯЛЬНОСТЬ (social NPS)	РАЗНООБРАЗИЕ (DIVERSITY)
ВОВЛЕЧЕННОСТЬ (ENGAGEMENT)	УДОВЛЕТВОРЕННОСТЬ (social CSI)	

### 8.5 Реализация интегрированных подходов

**8.5.0.1 Прямая интеграция** За последние 5 лет Андреа Колладон был единственным автором, который предложил измерять как сетевые, так и традиционные метрики и ввел индекс, который назвал «Semantic Brand Score» [130]. SBS сочетает в себе разнообразие, связность и вариант показателя «количество упоминаний», который Колладон называет «популярностью», чтобы обеспечить комплексную оценку важности бренда. Было показано, что SBS может иметь связь с некоторыми классическими бизнес-метриками, такими как число посетителей [132].

**8.5.0.2 Косвенная интеграция** В области прикладных исследовательских проектов сетевые и несетевые метрики могут использоваться вместе для комплексной оценки восприятия и позиции бренда. Например, отдельные агентства сочетают SBS, тональность, а также Social CSI и NPS. Андреа Колладон также использовал [131] SBS вместе с оценкой сходств имиджей бренда и тональностью.

### **8.5.1 Основные результаты**

- На сегодняшний день отсутствует систематическая научная дискуссия о показателях, используемых для оценки эффективности, восприятия и позиции бренда. Диалог в основном происходит на уровне отдельных акторов в сфере прикладных исследований.
- Традиционные метрики широко используются для описания эффективности и восприятия бренда. Даже самые распространенные из них (такие как охват или тональность) могут иметь несколько разные определения и формулы вычисления.
- Сетевые метрики в основном используются для оценки позиции бренда; однако семантические сети также могут быть источником информации о восприятии бренда.

#### **Будущие шаги**

1. Продолжать работу по систематизации метрик и показателей; разработать некий «золотой стандарт» интерпретации и операционализации метрик.
2. Расширять применение сетевого анализа для оценки позиции и восприятия бренда.

## **9 Здоровые и безопасные города: актуальные тренды исследований в научной литературе и социальных медиа**

### **9.1 Введение**

Статья представляет аналитический обзор по теме здоровых городов и городского здравоохранения. В основе обзора лежит методология сетевого анализа библиометрических источников — 5597 статей из Web of Science, отобранных по ключевым словам, и 2179 статей журнала Journal of Urban Health, являющегося флагманским в выбранной тематике. В дополнение проводится анализ публичного дискурса по теме здоровых городов на базе публикаций в социальных медиа, собранных компанией Brand Analytics. С содержательной точки зрения, проведенный анализ нацелен на выявление ключевых направлений исследований и обсуждений в области городского здоровья, а с методологической – демонстрирует, как разные виды анализа могут дополнять друг друга. В обзоре делается особый акцент на анализе обсуждений безбарьерной среды.

В центре концепции «здорового города» находятся люди, которые в нем живут, и которым, чтобы быть счастливыми, креативными и умными, нужно прежде всего быть здоровыми. Данная концепция признает, что здоровье является одним из ключевых ресурсов человека, наряду с материальными, социальными и другими возможностями. Само понятие «здоровье» здесь рассматривается шире, чем просто отсутствие болезней, и определяется как «состояние полного физического, душевного и социального благополучия» (согласно уставу ВОЗ).

Здоровье горожан зависит от множества городских характеристик, включая особенности социальной и физической среды, а также качество инфраструктуры. Со временем меняются представления, каким должен быть «здоровый город», а также приоритетность тех или иных характеристик. Для того, чтобы узнать, как трансформировалось поле изучения городского здравоохранения в научной литературе, а также понять, как научные исследования перекликаются

с дискурсивным пространством социальных медиа, мы провели анализ литературы с использованием алгоритмического подхода к анализу научных публикаций и применение различных приемов библиометрического анализа, которые позволяют выделять значимые работы и важных авторов на основе показателей их цитирования, а также делать выводы о развитии научных дисциплин на основе анализа библиографических сетей (цитирования, со-цитирования, библиографического сочленения, соприсутствия различных библиографических единиц, таких как авторы, журналы, ключевые слова и т.д.). В результате, можно делать выводы об актуальных трендах, обсуждаемых членами научного и экспертного сообщества.

Другим способом обзора трендов в различных тематических областях является анализ отражения этих тематик в публичном дискурсе, представленном данными социальных медиа. Для анализа данных такого типа используются алгоритмы, разработанные в количественном текстовом и сетевом анализе. В данном случае тематики и тренды формируются не экспертами, а широкой общественностью: наиболее часто встречающиеся тематики и связи между ними могут показывать всплески внимания к той или иной тематике, вызванные определенными причинами в общественном пространстве. Использование обеих тактик сбора и анализа данных позволяет сопоставить результаты и дать более детальное описание дискурса по теме.

Выявление актуальных тематик и трендов осуществляется на основе анализа данных научных публикаций и данных социальных медиа. Рассматриваемая база данных научных публикаций состоит из двух массивов публикаций, индексированных в БД Web of Science (WoS) — массива, собранного по поисковому запросу, основанному на ключевых словах по теме, а также массива публикаций из специализированного профильного журнала Journal of Urban Health (JoUH). Массив публикаций в социальных медиа состоит из упоминаний, собранных компанией Brand Analytics и содержащих эксперто отобранные ключевые слова по темам городского здравоохранения.

Мы предполагаем, что полученные результаты смогут быть восприняты читателями как руководство к тому, на какие публикации в первую очередь стоит обратить внимание при погружении в тематику здоровых городов, поэтому сопровождаем описание развернутыми таблицами в Приложении, размещенном на платформе GitHub<sup>18</sup>.

## 9.2 Обзор литературы

В этом разделе говорится про безопасность в широком смысле как ключевую характеристику здоровых городов. Здоровье города зависит от множества факторов, которые влияют на жизнеспособность и качество жизни его жителей: в т.ч. качество инфраструктуры и доступность услуг; качество окружающей среды; социальная структура и равенство; образование; экономическое развитие; социокультурный контекст; занятость населения.

Все эти факторы объединяет стремление людей чувствовать безопасность, социальную защищенность.

Чувство безопасности играет существенную роль в определении здоровья города. Когда жители города чувствуют себя безопасно, это может способствовать благополучию и улучшению их физического и психологического состояния.

---

<sup>18</sup>Приложение на GitHub: [https://github.com/Daria-Maltseva/Sociodigger/blob/main/UrbanHealth\\_Appendix.pdf](https://github.com/Daria-Maltseva/Sociodigger/blob/main/UrbanHealth_Appendix.pdf)

Налаживание безопасной среды способствует снижению риска физического насилия, преступлений и других опасностей. Когда жители ощущают безопасность на улицах и в общественных местах, они более склонны принимать участие в физической активности и здоровом образе жизни. Например, люди чаще будут гулять, заниматься спортом или выбирать альтернативы общественного транспорта, что способствует повышению уровня физической активности и снижению риска различных хронических заболеваний.

Чувство безопасности имеет прямое влияние и на психологическое благополучие горожан. Когда люди ощущают безопасность, у них меньше стресса, беспокойства и тревоги. Напротив, появление или усиление ощущения опасности может приводить к повышению уровня стресса, тревожности и депрессии у населения. Психические расстройства могут оказывать отрицательное влияние на жизнь горожан и увеличивать нагрузку на здравоохранение системы города.

Чувство безопасности способствует развитию здоровых социальных взаимодействий в городе. Когда люди чувствуют себя безопасно, они больше открыты для взаимодействий и участия в общественной жизни. У них есть доверие к своим соседям и другим горожанам, что способствует формированию поддерживающих и связывающих сообществ. Хорошие социальные связи и взаимодействие между людьми также могут иметь положительное влияние на здоровье в целом, поскольку они создают поддержку, снижают уровень стресса и повышают чувство принадлежности.

Ощущение безопасности в городе достигается не только благодаря отсутствию прямых физических угроз и низкому уровню преступности [223], [436]. Чувство безопасности в городе складывается из различных элементов и факторов, которые влияют на восприятие жителями уровня опасности и степени защищенности. В частности, признаки, ассоциируемые с небезопасностью: плохая видимость и слабая освещенность на улицах, низкая пешеходная и транспортная загруженность, отсутствие активных фасадов, слабо развитая сервисная инфраструктура и низкое количество многоквартирных домов [99].

Таким образом, чувство безопасности является важным аспектом при анализе здоровья города, так как оно оказывает влияние на психологическое и физическое благополучие жителей, социальные связи и включенность в общественную жизнь.

## 10 Методы сбора и обработки данных

В данной работе используется методология сетевого анализа, которая применяется для анализа двух типов данных, полученных из разных источников — научных публикаций из БД Web of Science и текстов из социальных медиа.

*Библиометрический сетевой анализ* является объединением методологии сетевого анализа с библиометрическим анализом, основанным на применении математических и статистических методов к данным научных публикаций, и позволяет изучать научные дисциплины и тематики как сети, состоящие из статей, авторов, журналов, ключевых слов и других связанных с ними библиографических сущностей. Применяемая в исследовании методология разработана В. Батагелем, А. Ферлигой и П. Дореаном [9] и уже использовалась для библиометрического сетевого анализа в различных областях: поля сетевого анализа [257] и направлений кластеринга и блокмоделинга [9], рецензирования [41], наукометрии и библиометрии [258], и др. В отечественной литературе описание этой методологии представлено в части анализа сетей цитирований между работами – в рамках описания алгоритмического подхода к отбору

источников для подготовки систематического обзора литературы [469], а также выделения актуальных тематик в социологии [451].

Сеть цитирования представляет собой граф, где вершины — это публикации, а ребра — отношения цитирования между работами. При анализе сети цитирований рассчитывается метрика входящей центральности, которая определяет наиболее цитируемые (т. е. значимые) научные работы. Выделение основных путей показывает цепочки из наиболее значимых работ в контексте цитирований в виде графа во времени, выделяемые с помощью алгоритма Search path count (SPC) [9], который для каждой конкретной связи в сети вычисляет индикатор веса проходов, или traversal weights [9]. Для проведения анализа сеть должна быть ациклична (не должна содержать «циклы», или последовательно связанные ребра, где один из узлов является и началом, и концом цепочки) и не содержать «петли» (ссылки на саму себя), иметь только одну связь между парами узлов, вес которой равен единице. Сеть с циклами трансформируется посредством алгоритма Preprint Transformation. Индикатор traversal weights отражает значимость ребра в сети, то есть вероятность того, что путь от искусственно созданного узла с наиболее поздним временем публикации (только цитирующего другие работы) к искусственно созданному узлу с наиболее поздним временем публикации (только цитируемому другими работами) будет проходить через данное ребро. К сети с рассчитанными весами применяются алгоритмы выделения основного пути (Main path) или ключевых путей (Key routes). При применении алгоритма основного пути (Main path) для каждой публикации «верхнего уровня» (имеющей ссылки на другие работы, но не цитируемой другими) на основе последовательного выбора вершин сети по определенному правилу конструируется цепочка, ведущая к публикациям «нижнего уровня» (цитируемым другими работами, но не имеющим цитирований). Цепочка с максимальным значением показателя отбирается в качестве основного пути. При применении алгоритма поиска ключевых путей (Key Routes) в сети выделяется не один, а несколько возможных путей — через увеличение количества включенных в основной путь связей.

В основные пути входят не только наиболее цитируемые и значимые работы, но те работы, которые ссылаются на значимые в предыдущем временном отрезке публикации. Другой подход, позволяющий определить локально важные участки сети, где узлы сильнее связаны друг с другом, чем с узлами в остальной части сети — Islands approach, или подход островов [9], — был использован для выделения интересных для изучения подгрупп в сети. При использовании этого подхода указываются границы размера подгрупп в сети, которые могут быть найдены (например, от 2 до 100 узлов).

Определяя тематическую направленность выделенных в ходе анализа сети цитирований публикаций, можно сформировать представление об основных представленных в научном дискурсе тематиках, а также определить их путь развития и преемственности в той или иной предметной области. Анализ наиболее цитируемых работ и работ, вошедших в основные пути, стал базой выделения и изучения актуальных тематик исследований в рассматриваемой нами предметной области городского здоровья.

Помимо анализа сети цитирований, для анализа была использована двумодальная сеть работ и ключевых слов (в этой сети узлы одной части сети (работы) связаны с узлами другой части сети (ключевые слова), но не связаны друг с другом; вес каждой связи равен единице.), на основе которой через перемножение была построена сеть со-встречаемости ключевых слов в описании одной статьи. Анализ сетей ключевых слов подразумевал как уже упомянутый расчет степени входящей центральности (наиболее часто встречающиеся в публикациях ключевые слова), так и выделение подгрупп в сети

с помощью подхода островов, позволяющего выявлять локально важные группы узлов, более тесно связанные друг с другом, чем с остальными узлами сети.

Анализ данных библиометрического сетевого анализа выполнен в программе Rajek для анализа и визуализации больших сетей [289], а также в программе R.<sup>19</sup> Документация по анализу данных представлена в открытом доступе по ссылке.<sup>20</sup>

Анализ социальных медиа представляет широкий класс методов анализа, применяемых к данным из социальных сетей, форумов и других интернет-площадок, на которых общаются пользователи [432]. Для выделения тематик обсуждений в социальных медиа в данном исследовании используется метод тематического моделирования. Latent Dirichlet Allocation (LDA) — подход, используемый в тематическом моделировании на основе вероятностных векторов слов, которые указывают на их релевантность текстовому корпусу [48].

Для поиска сетевой структуры обсуждений используется сетевой подход к анализу текстов, который включает в себя создание «ментальной модели» — сети отношений между понятиями, встречающимися в текстах [61].

Анализ данных социальных медиа выполнен с помощью языка программирования Python (пакеты nltk<sup>21</sup> и networkx<sup>22</sup>, визуализация сети — в программе для визуализации сетевых данных Gephi<sup>23</sup>.

## 11 База данных

### 11.0.1 Библиометрические данные

Стратегии сбора данных для библиометрического анализа подразумевают поиск публикаций по ключевым словам в базах данных научных публикаций, таких как Web of Science, Scopus и др. [41], отдельным журналам [184] или коллекциям публикаций по определенной тематике, подготовленных экспертным сообществом. В настоящем исследовании данные для библиометрического анализа были получены из международной БД научного цитирования Web of Science (Core Collection) с помощью двух стратегий сбора данных. Именно эта база использовалась для сбора в связи с тем, что входящие в нее работы имеют полные библиографические описания, включающие списки цитируемых источников (поле CR в описании), наличие которых является важным для построения сети цитирований. Входящая в WoS база отечественных публикаций, входящих в РИНЦ (RSCI, Russian Science Citation Index) такой информации, не содержит, поэтому ее нельзя было использовать в качестве источника информации.

- Первый массив был получен с помощью поисковых запросов по ключевым словам на английском языке, отобранным экспертыным образом. Выдача результатов по следующим ключевым словам составила: «urban health» - 5508 статей, «health-saving environment» – 4 статьи, «barrier-free environment» – 85 статей; общее количество статей в массиве по поисковым запросам составило 5597 статей. Для поиска рассматривались также ключевые «public health», «mental health» и «medical

<sup>19</sup>\*R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: <https://www.R-project.org/>. (accessed: 19.12.2022).

<sup>20</sup>\*GitHub: <https://github.com/Daria-Maltseva/Sociodigger/wiki/UrbanHealth>

<sup>21</sup>\*NLTK Project // Natural Language Toolkit / NLTK - Developers and contributors (by Steven Bird, Edward Loper, Ewan Klein and others). URL: <https://www.nltk.org> (accessed: 19.12.2022).

<sup>22</sup>\*NetworkX // Software library. (by Aric Hagberg Pieter Swart Dan Schu and others). URL: <https://networkx.org/> (accessed: 19.12.2022).

<sup>23</sup>\*R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: <https://www.R-project.org/>. (accessed: 19.12.2022).

prevention\*», однако в связи с их чрезмерно широким смыслом, выходящим за рамки городского здоровья, их использование без добавления уточняющих слов привело бы к слишком большому количеству статей в выдаче результатов, и, что более важно, к смешению результатов анализа на более широкую тему. Использование этих ключевых слов вместе с указанными более узкими ключевыми словами привело к 1920 статьям в выдаче, все из которых входили в полученную выборку из 5597 статей.

- Второй массив был получен путем сбора всех статей, опубликованных в журнале *Journal of Urban Health*, посвященном вопросам общественного здравоохранения в городских районах и благополучия жителей городов, который можно считать флагманским для данной области исследований. Журнал выпускается Springer Science+Business Media и New York Academy of Medicine; его история ведется с 1851 года, а импакт-фактор в 2021 году составляет 5.801. Общее количество статей в массиве по журналу составило 2178 публикаций.

Второй массив публикаций по журналу полностью входит в первый массив публикаций, найденных по ключевым словам. Несмотря на это, было принято решение проанализировать оба массива, чтобы сравнить две стратегии сбора данных.

Для краткости обозначения сети цитирований, построенные из массивов по ключевым словам и журналу, называются ниже **CiteSQ** и **CiteJ**. Были сформированы отдельные массивы для публикаций с полным описанием (хиты) и публикации, появившиеся только среди процитированных источников (только цитируемые работы). Количество хитов равняется количеству работ, найденных при поиске в базе данных (за вычетом дублей) и составляет 5590 и 2177 публикаций в двух сетях соответственно, а количество цитируемых ими работ составляет 144152 и 54126, соответственно (с учетом чистки данных от анонимных работ).

### 11.0.2 Данные социальных медиа

Данные для анализа социальных медиа были собраны с помощью системы мониторинга и анализа социальных медиа и СМИ Brand Analytics<sup>24</sup>, которая охватывает социальные сети ВКонтакте, Одноклассники Instagram, Facebook<sup>25</sup>, YouTube, TikTok, Twitter и др., а также блоги, форумы, отзывы, маркетплейсы, сервисы карт, сторы мобильных приложений, публичные каналы и чаты Telegram, онлайн СМИ, сайты госучреждений, рынкообразующих компаний и организаций. С учетом тематической направленности данного выпуска и статей, которые в нем опубликованы, а также общей информационно-новостной повестки, экспертоно были отобраны 6 тем, связанных со здоровьем города: безбарьерная среда (1,41 млн упоминаний), мусорные полигоны (2,24 млн упоминаний), семейное насилие (2,03 млн упоминаний), высотная застройка (772 тыс. упоминаний), спортивные площадки (1,42 млн упоминаний) и эко-привычки горожан (909 тыс. упоминаний). В качестве темы для более детального анализа были выбраны вопросы, связанные с безбарьерной средой в городе. Подробная документация принципов отбора публикаций в социальных медиа, включая использованные поисковые запросы и стоп-слова, представлена в Приложении на GitHub (Приложение 2<sup>26</sup>).

<sup>24</sup> <https://br-analytics.ru>

<sup>25</sup> Facebook и Instagram принадлежат компании Meta Platforms Inc., признанной экстремистской организацией, чья деятельность запрещена на территории РФ

<sup>26</sup> Приложение на GitHub: [https://github.com/Daria-Maltseva/Sociodigger/blob/main/UrbanHealth\\_Appendix.pdf](https://github.com/Daria-Maltseva/Sociodigger/blob/main/UrbanHealth_Appendix.pdf)

Данные представляют собой текстовые упоминания пользователей социальных сетей о безбарьерной среде в городе — мнения, суждения, опыт взаимодействия. В дополнение к текстовым данным, система Brand Analytics предоставляет метаинформацию о площадках, на которой размещены публикации, об авторах, тональности публикаций и пр.

## 11.1 Результаты

### 11.1.1 Результаты анализа библиометрических данных

**11.1.1.1 Анализ цитирований** Ниже представлено распределение количества публикаций по годам выхода по двум анализируемым массивам для хитов — работ, найденных по поисковому запросу в БД WoS (Рис. 1) и всех публикаций, включающих хиты и цитируемые ими работы (Рис. 30). В первом случае распределение приведено с 1990 г., во втором — с 1950 г., в связи с незначительной представленностью более ранних работ в массивах (отметим, что Journal of Urban Health ведет свою историю с 1851 г., а WoS позиционируется как БД, индексирующая информацию о публикациях с 1900 г.). Рис.31 показывает, что количество публикаций, отвечающих заданному нами поисковому запросу, плавно возрастает с 2005 г. — от 150 статей до 500 в год; количество публикаций в JoUH держится на примерно одинаковом уровне (что объясняется ограниченными публикационными возможностями журнала). Рост числа релевантных публикаций наблюдается и по графику распределения всех публикаций (хитов и цитируемых ими работ, Рис.31). Вместе с тем, видно, что в массиве по журналу пик публикаций приходится на 2000-е гг., тогда как в массиве по поисковому запросу идет плавное увеличение числа публикаций, достигающее максимума в 2013–2014 гг. Уменьшение числа публикаций в последние несколько лет является нормальным в библиографических исследованиях и объясняется публикационным циклом и закономерностями в цитировании (поздние работы цитируют более ранние работы).

Количество публикаций по годам в сети **CiteSQ** (Urban Health) и **CiteJ** (JoUH) — хиты

Figure 30: Количество публикаций по годам в сети **CiteSQ** (Urban Health) и **CiteJ** (JoUH) — хиты

Количество публикаций по годам в сети **CiteSQ** (Urban Health) и **CiteJ** (JoUH) — все работы

Figure 31: Количество публикаций по годам в сети **CiteSQ** (Urban Health) и **CiteJ** (JoUH) — все работы

Были составлены списки работ, входящих в топ-50 наиболее цитируемых работ в сети цитирований CiteSQ<sup>27</sup>. Работой с наибольшим количеством цитирований является статья Сэмпсона и коллег (Sampson, Raudenbush, & Earls, 1997) в журнале Science 1997 года, посвященная связи сплоченности в соседских сообществах и уменьшению насилия. Три следующие работы имеют примерно одинаковое количество цитирований (72, 71 и 70). Статью Galea & Vlahov 2005 года можно назвать программной, поскольку в ней делается обзор различных исследований влияния жизни в городах на здоровье населения и говорится о важности изучения городского здоровья как отдельной области исследований, с такими широкими

<sup>27</sup>В список топ-50 наиболее цитируемых публикаций попали работы формата «WORLD\_H(2022)», которые сложно идентифицировать в связи с тем, что используемое название может обозначать разные работы, например: World Bank, 2022, LIFE EXPECTANCY BIRT, World Bank, 2022, MORTALITY RATE UNDER, World Bank, 2022, HOSP BEDS PER 1000 P, World Bank, 2022, MATERNAL MORTALITY R. В связи с таким форматом программа WoS2Pajek считывает первое слово в названии организации как фамилию автора, выбирает из второго слова первый знак как инициал его имени, указывает в скобках год публикации и не указывает после никаких данных, т. к. информации о выпуске журнала и странице публикации в этих описаниях нет. В связи с тем, что под короткое описание может подходить несколько публикаций, показатель его цитирования равен сумме цитирований разных работ. Более корректным является исключение таких работ из списков (что было сделано в обоих массивах).

темами как физическая, социальная среда и доступ к медицинским и социальным услугам, а также о методологических и концептуальных проблемах. Работа Williams & Collins 2001 года посвящена расовой сегрегации по месту жительства, высокой для афроамериканцев в США, которая влечет различия в социально-экономическом статусе, влекущие расовые различия в состоянии здоровья. Работа Radloff, опубликованная значительно раньше — в 1977 году, — посвящена описанию шкалы для измерения депрессивной симптоматики среди населения.

Большинство работ с высокими показателями цитирований также носит программный, общий характер и посвящены связи здоровья с урбанизацией и жизнью в городах в целом (18, 33), соседскими общиными и сообществами (10, 13, 21, 40) и социальными условиями как причинами заболеваемости (23, 41), публичными пространствами (50), зелеными насаждениями (34), объектами физической активности (46), распространением ВИЧ (45), жизнью в трущобах (44), а также посвящены вопросам улучшения условий жизни в городах для снижения заболеваемости и различий в состоянии здоровья (8, 17, 30), планированию здоровой городской среды (9, 38), в т.ч. в развивающихся странах (11, 39). В нескольких работах описываются вопросы, актуальные для формирования городской среды в США — политика апартеида, приведшая к сегрегации по расовой принадлежности (36), а также политика городского планирования, приведшая к упадку многих городских кварталов (35). Несколько представленных работ посвящены вопросам операционализации, концептуализации и измерения влияния городской среды на здоровье (20, 37), а также общим методологическим вопросам — выборке, направляемой респондентами (11, 48) и целевой выборке (32).

Аналогичные результаты анализа по сети **CiteJ** приведены в Таблице 5. Тогда как наибольшее количество цитирований в этой сети имеет работа Heckathorn 1997 года, посвященная описанию выборки, движимой респондентами, для изучения скрытых популяций, три следующие за ней работы также входили в топ публикаций по Таблице 5: тема сплоченности в городских сообществах и насилия, шкала депрессивности и расовая сегрегация по месту жительства. В целом, 21 из 47 представленных в Таблице 6 работ присутствует также в Таблице 5 (пересекающиеся в двух таблицах статьи выделены серым цветом). Значительное количество не представленных в Таблице 5 работ посвящено вопросам здоровья маргинальных групп в городах — распространения ВИЧ и других инфекций через обмен шприцами среди людей, употребляющих наркотики, заболеваний, передающихся половым путем, здоровья бывших заключенных, — а также психологических проблем и депрессии. Можно выделить работы, посвященные физической активности в городских пространствах (24, 40, 46) и избыточного или недостаточного веса детей (31), а также методологии проведения исследований (27, 49). 21 публикация вошла в оба массива.

Полные таблицы представлены в Приложении на GitHub по указанной ранее ссылке.

На Рис. 32 приведен основной путь в сети **CiteSQ**, состоящий из 12 работ. По 3 работы опубликованы в журналах *Journal of Urban Health* и *Health Policy and Planning*, по 2 работы — в журналах *Environment and Urbanization* и *Lancet*. Наиболее ранняя работа (Ludwick) опубликована в 1998 году и посвящена влиянию курения на преждевременную смертность жителей в Гарлеме, одном из районов Нью-Йорка. Следующая статья тоже же автора (2000) уже носит программный характер и, основываясь на данных эмпирических исследований, говорит о важности изучения структурных факторов, влияющих на состояние здоровья и смертность в районах современных городских гетто и проживания меньшинств, которые должны приниматься во внимание специалистами в области городского здравоохранения. Три следующие работы авторов Vlahov и Galea (2002, 2003, 2005) обозначают направления городского

здравоохранения как самостоятельной дисциплины, опирающейся на наработки в области экологии, эпидемиологии и социологии. Влияющие на здоровье факторы рассматриваются в рамках трех широких тем — социальной среды, физической среды и доступа к медицинским и социальным услугам. Работы 2002 и 2005 годов также были выделены в ходе анализа наиболее цитируемых публикаций в сети CiteSQ. Следующие статьи посвящены стратегиям управления здоровым городом (2007), а также вопросам достижения справедливости в отношении здоровья в городах (2007, 2008). Далее в 2011 и с большим отрывом — в 2017 — годах в основном пути находятся работы, поднимающие вопросы состояния здоровья жителей трущоб. Наконец, четыре наиболее недавние статьи (2017–2022) посвящены обсуждению роли моделей медицинского обслуживания в городских сообществах (*urban community health workers*) в решении проблем неравенства в состоянии здоровья в городах.

Таким образом, в работах основного пути от понимания важности учета социальных факторов на основе эмпирических исследований тематика переходит к формированию городского здравоохранения как самостоятельной междисциплинарной области исследований и вопросам достижения справедливости в отношении здоровья, недостижимой для жителей городских трущоб; после чего обсуждаются попытки практического решения этих проблем через медицинское обслуживание в городских сообществах.

Основной путь (слева) и ключевые пути (справа) в сети CiteSQ

Figure 32: Основной путь (слева) и ключевые пути (справа) в сети CiteSQ

Тематика, представленная в основном пути, может быть описана подробнее посредством анализа работ, входящих в ключевые пути (Таблица 1 в Приложении на GitHub, где работы, входящие в основной путь, отмечены кодом 1, а публикации из ключевых путей — кодом 2; наличие двух кодов означает, что публикации входят в обе структуры). Сюда входит 27 работ, из которых 12 включены в основной путь, описанный ранее. Ранние работы (1998–2000 гг.) в левой нижней части рисунка (Speers, Leviton, Freudenberg) в этой структуре посвящены предотвращению болезней и пропаганде здорового образа жизни (что подразумевает предотвращение сердечных заболеваний, использования наркотиков, насилия, передачи ВИЧ-инфекции). Публикации (2001–2002) в правой нижней части рисунка (Eisinger, Higgins и далее) посвящены формированию исследовательских центров на базе сообществ в различных городских условиях. Работы в средней части рисунка значительно пересекаются с работами из основного пути и описывают тематику городского здоровья в связи с социальным неравенством, проживания в трущобах и — в более поздних работах — влияния медицинского обслуживания для решения этих проблем. Последние публикации (2021–2022) в правой верхней части рисунка (Dachaga и др.) посвящены безопасности землевладения (Land Tenure Security). Таким образом, дополнительные тематики, найденные в ключевых путях, относятся к пропаганде здорового образа жизни и тематике безопасности землевладения, а также более подобно говорят о формировании исследовательских центров на базе сообществ для решения проблем здравоохранения.

Основной путь в сети **CiteJ**, состоящий из 8 работ в журнале *Journal of Urban Health*, показан на Рис. 33 (слева). Он начинается с двух работ 2001 года, посвященных влиянию тюрем и исправительных учреждений на здоровье населения. Три следующие работы (2001, 2005 и 2006 гг.) посвящены вопросам реинтеграции выходцев из исправительных учреждений в городские сообщества и их влиянию на общественное здоровье, в т.ч. с помощью специализированных социальных программ (рассматривается их влияние на повторные аресты, использование наркотиков и ВИЧ). 2010 года фокус работ смешается

на рискованное сексуальное поведение мужчин, употребляющих наркотики / выходящих из тюрьмы / практикующих сексуальные связи с мужчинами. Таким образом, в этом основном пути превалирует тематика здоровья группы населения, связанной с тюремным заключением, и ее влияние на общественное здоровье в целом.

Основной путь (слева) и ключевые пути (справа) в сети CiteJoUH

Figure 33: Основной путь (слева) и ключевые пути (справа) в сети CiteJoUH

Если анализировать работы во втором полученном основном пути подробнее (Таблица 2 в Приложении 1), стоит отметить, что в него вошло 23 статьи, из которых 8 также входят в первый основной путь, описанный выше. Как видно на Рис. 33, ключевые пути разбиваются на три части. Структура справа, по сути, повторяет основной путь. Дополнительные работы посередине рисунка также связаны с маргинальными группами населения и посвящены вопросам безрецептурной продажи шприцев в контексте предотвращения ВИЧ (2000–2013). Структура слева на рисунке представляет тематику связи урбанизации и здоровья и становления городского здравоохранения как самостоятельной дисциплины (выделенная в анализе сети **CiteSQ**), а работы 2008 и 2020 гг. посвящены материнскому и детскому здоровью.

В каждой сети цитирований были выделены острова — плотно связанные друг с другом узлы, отражающие локальное важные участки сети. В сети **CiteSQ** основной остров состоит из 98 узлов и включает 26 из 27 работ, входящих в структуру основного и ключевых путей (эти работы отмечены кодом 3 в Таблице 1 в Приложении 1). Кроме отмеченных выше, первые по времени появления публикации в этом острове охватывают тематики проблем общественного здравоохранения в развивающихся странах и семьях с малым достатком, социальному доверию и социальным связям в связи со здоровьем, употреблению наркотиков и связанной с ними смертностью. С середины 2000-х достаточно много публикаций посвящены связи здоровья с жизнью в городе, соседскими общинами, жизнью в трущобах, а также развитию городской политики и планированию в области здоровья и справедливости в отношении здоровья; встречаются работы по построению индекса здоровья населения, а также работы, посвященные распространению COVID-19.

В сети **CiteJ** основной остров состоит из 92 узлов и включает 14 из 23 работ, входящих в структуру основного и ключевых путей (эти работы отмечены кодом 3 в Таблице 2 в Приложении 1). Дополнительные темы, проявляющиеся в публикациях в этом острове, не обозначенные в ключевых путях в этой сети, посвящены тематике исследовательских центров на базе сообществ (встреченных в ключевых путях в сети **CiteSQ**), употребления наркотиков в более широком контексте, чем встречалось в ключевых путях в этой сети, репродуктивного здоровья женщин и здоровья детей (в т. ч. заболевания астмы).

Интерес представляют работы, вошедшие в острова с меньшим количеством узлов. Таблицы 3 и 4 в Приложении 1 показывают состав выделенных кластеров («островов») для каждой сети цитирований.

В сети **CiteSQ** можно выделить тематики контроля и выявления различных заболеваний, встречающихся у жителей городских агломераций: контроля астмы в городских пространствах в рамках специализированных проектов (кластер 3) (кластеры 2 и 5 были исключены из описания ввиду их небольшого размера), профилактики рака с помощью маммографических скринингов (10), гепатита В и ВИЧ (11), выявления миомы матки (16). Сюда же можно отнести обсуждение профилактического подхода к здоровью старению пожилых людей (12). Другая группа тем посвящена здоровью бездомных

(18) и заключенных (19), при этом последний кластер тематически совпадает с основным путем, полученным в сети **CiteJ**; в контексте бедности и бездомности обсуждаются также азартные игры (9). Два кластера рассматривают тематику домашнего насилия и первичной медицинской помощи (4), использования оружия и убийств (8). Группа кластеров посвящена темам, связанным с употреблением наркотиков: оценке потребителей инъекционных наркотиков в крупных мегаполисах (13), оценке программы продажи стерильных шприцев (14), связи потребления наркотиков с распространением ВИЧ (6) и выселением из жилых помещений (15). Сюда же можно отнести тематики, посвященные рискованному сексуальному поведению мужчин, занимающихся сексом с другими мужчинами (7) и состоянию здоровья представителей сексуальных и гендерных меньшинств, поскольку в них также поднимается тематика заболевания ВИЧ (17). Тематика (20) посвящена очень специфической теме — существованию городских крыс в мегаполисах.

В сети **CiteJ** ряд тем посвящен оценке различных аспектов городской среды и ее влиянию на физическое состояние горожан: физической активности в окружающей искусственной и социальной городской среде (4), городским пространствам, озеленению, использованию зонирования и точкам продажи алкоголя (15), городской пищевой среде и индекс массы тела, ожирение (9). Также выделяется тематика профилактики — контроль астмы (11), профилактический педиатрический уход за маленькими детьми (8). К профилактике можно отнести модель укрепления здоровья стареющего населения (14), здоровое старение в городах, вовлечение пожилых людей в волонтерскую деятельность для укрепления здоровья (20).

Выявляются темы про расовые различия в состоянии здоровья в городских сообществах (10), самооценки состояния здоровья, в т. ч. в контексте сегрегации и джентрификации (13). Связанными с ними являются темы здоровья бездомных людей в контексте изменения климата (17) и жилищной нестабильности (18). Как и в сети выше, представлена также тематика лицензирования продажи и незаконного оборота огнестрельного оружия, убийства в городских округах (6).

Как и в описанных выше результатах, в данной сети присутствуют темы, связанные с использованием наркотиков, в том числе смертность от передозировки наркотиками (3), рискованные сексуальные отношения (МСМ), использование наркотиков, заболевания, передающиеся половым путем (5), распространение гепатита В, ВИЧ у потребителей инъекционных наркотиков, программы вакцинации (12). Тематика ВИЧ представлены через расовые и этнические различия в распространенности и рисках ВИЧ (19) и профилактику ВИЧ для женщин (16). Сюда же тематически ложится тематика заболеваний, передающихся половым путем и рискованного сексуального поведения (МСМ) у заключенных и выходцев из мест лишения свободы, которая напоминает основной путь, полученный по сети **CiteJ**. Отличительной особенностью анализа данной сети является получение кластера, в котором обсуждаются методологические вопросы исследований в области городского здравоохранения — оценка выборки, используемой для изучения труднодоступных групп (целевая выборка, выборка, управляемая респондентом) (7).

**11.1.1.2 Анализ ключевых слов** Для анализа ключевых слов используется двумодальная сеть работ и ключевых слов **WK**, полученная в результате трансформации исходных данных массива, полученного по поисковому запросу, в программе **WoS2Pajek**, которая состоит из 149740 публикаций и 8254 ключевых слов.

Ввиду малого количества работ с выделенными ключевыми словами до 1990 года, внимание к различиям в частоте появления ключевых слов в работах было направлено к публикациям, опубликованным не ранее 1990 года. Публикации были разделены на 4 периода по десятилетиям: 1990–2000, 2001–2010, 2011–2020 и 2021 и далее. Для каждого периода была построена редуцированная версия сети работ и ключевых слов и подсчитана метрика входящей центральности indegree, позволяющая оценить, насколько часто в указанном периоде конкретное слово появлялось в списке ключевых слов.

Естественным образом, наиболее часто встречающимися, вне зависимости от временного периода, являются слова «health» и «urban» как определяющие для дисциплины. Тем не менее, в распределении прочих позиций наблюдается ощущимая вариация, передающая изменения в значимых инфоповодах каждого десятилетия. Наиболее наглядно изменения в темах академических публикаций для обозначенных четырех временных периодов демонстрируют «облака слов» (Рис. 34 – 35), которые позволяют получить своеобразный «срез» академических инфоповодов за десятилетие (ключевые слова «health» и «urban» исключены из визуализаций, чтобы подчеркнуть существующие различия).

Облака слов для периода 1990–2000 гг. (слева) и 2001–2010 гг. (справа)

Figure 34: Облака слов для периода 1990–2000 гг. (слева) и 2001–2010 гг. (справа)

Облака слов для периода 2011–2020 гг. (слева) и 2021-гг. (справа)

Figure 35: Облака слов для периода 2011–2020 гг. (слева) и 2021-гг. (справа)

1990-е и «нулевые» годы отмечены ростом обеспокоенности распространением ВИЧ. В публикациях часто встречаются понятия риска и отсылки к организованным коллективным формам здравоохранения. Отдельно следует отметить академический фокус на наркотических веществах как факторе распространения ВИЧ. Новое десятилетие обозначило иные тренды, в первую очередь, социальный аспект здравоохранения, и на передний план выходят концепты коммунальности. Гораздо чаще встречаются понятия социального, сообществ и соседств, а также публичности и неравенства. Этот тренд частично продолжается в начале 2020-х, однако, ожидаемо, наиболее релевантной и популярной проблемной областью первых лет 20-х годов стала пандемия COVID-19. Примечательно, что пандемия перекрыла, в том числе, и гендерный вопрос: впервые за 30 лет ключевое слово «женщина» не вошло в топ-20 самых часто встречающихся (Рис. 36). Резкие падения интереса к гендерным аспектам здравоохранения фиксируются в конце 2010-х, однако именно 2021 год сводит ранее достаточно устойчивый рост практически к нулю.

Распределение частоты употребления ключевого слова «woman»

Figure 36: Распределение частоты употребления ключевого слова «woman»

Наблюдая распределение частот упоминания наиболее популярных за весь исследуемый диапазон ключевых слов (Рис. 37), можно отметить, что многие из них являются как раз отражением временно релевантного инфоповода или академической моды. Так, тематика ВИЧ имеет два острых пика активного появления в академических работах, однако исчезает по мере сокращения медиа-покрытия и экстренности эпидемии в развитых странах, равно как и тема наркотических веществ. Примечательно распределение частот для ключевого слова «factor»: устойчивый рост сменяется резким падением и еще более резким новым пиком. Поскольку поиск факторов, как правило, привязан к крупным проблемам, наподобие

эпидемий, он распространяется по мере развития проблемы и угасает с её разрешением в ожидании новой. Неоднородна также динамика использования слов «public» и «state». Эти слова используются в среднем достаточно редко, однако имеют яркие пики в период 2010-х годов. Выше была обозначена «социальная» направленность академических текстов 2010-х, и популярность проблематики публичности или государственных мер отражает этот тренд.

Динамика частоты упоминаемости ключевых слов «HIV», «factor», «state» и «public»

Figure 37: Динамика частоты упоминаемости ключевых слов «HIV», «factor», «state» и «public»

Сеть соприсутствия ключевых слов по всем годам состоит из 8254 ключевых слов. Построенная сеть нормализована с использованием фракционного подхода [37]. С помощью подхода островов был выделен основной кластер связанных друг с другом ключевых слов, представленный на Рис. 38.

Сеть со-встречаемости ключевых слов ККп по всем годам

Figure 38: Сеть со-встречаемости ключевых слов ККп по всем годам

Так, выявляются три ключевых направления проблемных областей. В первую очередь можно выделить тесно внутренне связанный кластер слов, соприсутствующих с «hiv». Этот блок отражает развитие тематики ВИЧ и затрагивает сопряженные с ней демографические и социальные аспекты (такие как употребление наркотических веществ, сексуальное поведение и проч.). Вторым суб-кластером сети является организация здравоохранения (блок слов, соприсутствующих с «care»). Сюда входят слова, относящиеся к управлению и организационным аспектам медицины («management», «service» и т. п.). Далее, можно выделить небольшую группу связанных слов, где наиболее связанным с остальной сетью является «community», часто используемое для описания партиципаторных практик локальных сообществ, связанных с медицинским уходом. Наконец, менее четко выраженный, но когерентный в смысловом отношении блок соприсутствующих слов ассоциируется с проблемой распространения обиходных практик здорового образа жизни горожан («physical environment», «activity», «build»). Другими словами, здесь ярче прочего видна связка медицины с городом, как средой обитания, в которой предстоит осуществлять как можно более здоровую жизнедеятельность большинству жителей современных развитых стран. Здесь выделяются такие выражения как «green space», «planning», отражающие распространенный фокус на трансформации городской среды с целью оздоровления её жителей. Отдельно следовало бы подчеркнуть малочисленные, но выделяющиеся на фоне остальной сети слова, посвященные роли государства как актора в оздоровлении городской среды. В отличие от риторики «лечения» или «заботы» о пациентах, государственный дискурс включает не «сообщества» и не «пациентов», но «единицы населения» («population unit»). Связь со словом «surveillance» («надзор») подтверждает биополитический фокус рассмотрения роли государства в публикациях по данной тематике.

Несмотря на подчеркнутые выше смысловые различия в сети соприсутствия ключевых слов, важно отметить, что сами кластеры имеют связи друг с другом. Возможны и популярны преломления дискуссий о ВИЧ с перспективы организации здравоохранения. Лечение предполагает академический и социальный контекст, предоставляющий соучастие для эффективного достижения целей. В свою очередь, специфику как образования сообществ, так и предоставления медицинской помощи имеет городская среда.

В редуцированной сети работ и ключевых слов за 1990–2000 гг. число работ составило 26519, а ключевых слов — 1603, в сети 2001–2010 гг. — 51316 работ и 3778 ключевых слов, в сети 2011–2020

гг. — 55535 работ и 6195 ключевых слов, а в сети 2021–2022 гг. — 2659 работ и 3034 ключевых слов. Редуцированные версии сетей работ и ключевых слов были использованы для построения сетей со-встречаемости ключевых слов по 4 указанным периодам (число узлов в них равняется количеству ключевых слов в каждом периоде). В каждой сети с помощью подхода островов также были выделены основные кластеры связанных ключевых слов (Рис. 11.1.1.2 - 40). Сеть со-встречаемости ключевых слов KKн1 за 1990–2000 гг.

В сети, отражающей первый временной период (Рис. 11.1.1.2), наблюдается подчеркнуто управляемческий академический дискурс. Основными темами публикаций являются либо новые методы и результаты академических исследований, либо организационные вопросы лечения. Эти два направления достаточно ощутимо разъединены и соединяются только в точках, отражающих массовые болезни, где заболевание перестает быть сугубо объектом исследования и становится предметом политических стратегий (см. «tuberculosis», «hiv»).

Сеть со-встречаемости ключевых слов KKн2 за 2001–2010 гг.

Figure 39: Сеть со-встречаемости ключевых слов KKн2 за 2001–2010 гг.

В период 2001–2010 годов (Рис. 39) дискурс о «здравой» городской среде центрируется вокруг единой проблемы ВИЧ-инфекций. Сеть также наглядно показывает основную гипотезу о распространении, связанную с сексуальным поведением и употреблением наркотических веществ. Вероятно, под эгидой заболевания, характеристики распространения которого не могут быть проанализированы в отрыве от социального контекста, возник дискурс о социальном неравенстве и роли сообществ в медицине. Настоящая сеть имеет тематические «отростки», связывающие сообщества с академическими исследованиями, которые включают расовую и гендерную справедливость в систему координат организации медицины. Наиболее необычна, однако, связь, образующаяся у центрального набора слов с понятиями социальной ответственности и биоэтики через категорию риска. Так, именно риск оказывается связующим звеном между медицинской практикой и социальной проблематикой, актуализируя превенцию, нежели чем лечение.

Сеть со-встречаемости ключевых слов KKн3 за 2011–2020 гг.

Во втором десятилетии 21 века вопрос ВИЧ отошел на второй план, однако «риск» сохранил свое центральное положение в дискурсе. Маркировка «риска» удобно локализует болезнь в точке пересечения определенных социальных факторов, переводя внимание с уже больных на тех, кто мог бы потенциально заболеть. Расширяется географический спектр городского, к которому обращаются исследователи: если в предшествующие периоды слово «город» («city») как имя собственное проявляло себя лишь в сочетании «New York City», то на Рис. 12 уже видны отсылки к европейским городам или Гонконгу. Выделяются три смежные проблемные области, уже близко сопоставимые с наблюдаемыми в общей сети (Рис. 38): факторы риска, городская среда и организация здравоохранения.

Сеть со-встречаемости ключевых слов KKн4 за 2021–2022 гг.

Figure 40: Сеть со-встречаемости ключевых слов KKн4 за 2021–2022 гг.

В сети за 2021–2022 гг. доминирует пандемия COVID-19. Сеть соприсутствующих ключевых слов (Рис. 40) иллюстрирует изобилие социальных, экономических и пространственных измерений пандемии. Тема вакцинации оказывается сравнительно исключенной, в то время как на первый план

обсуждений выходит сравнительный анализ эффективности реакции на возникшую угрозу между странами, особенности городских и сельских возможностей предохранения от болезни и меры, которые можно принять, чтобы обезопасить публичное пространство. Особенности этой сети наиболее ярко подчеркивают произошедший с 1990-х годов сдвиг в медикалистском дискурсе в пользу фокуса на публичности: главными героями статей о медицине становятся не сами вирусы и их изучение, но сообщества, города, государства — иными словами, социальные единицы, структура которых зачастую играет решающую роль в определении развития болезни.

### **11.1.2 \*Результаты анализа данных социальных медиа**

Наиболее крупными темами для обсуждения за год стали мусорные полигоны (2,24 млн упоминаний) и семейное насилие (2,03 млн упоминаний). Следом идут спортивные площадки (1,42 млн упоминаний) и безбарьерная среда (1,41 млн упоминаний). Менее заметные на фоне остальных — эко-привычки горожан (909 тыс. упоминаний) и высотная застройка (772 тыс. упоминаний).

В динамике обсуждений тем видна сезонность и реакция интернет-пользователей на новостные события. В марте 2022 все темы, кроме вопросов семейного насилия, отошли на второй план из-за превалирования в обсуждениях специальной военной операции России в Украине и ее последствий. Обсуждение домашнего насилия продолжилось на том же уровне и в следующий месяц даже немного увеличилось на фоне общей тревожности общества и изменения информационной повестки. Тема мусорных полигонов особо активно обсуждалась в октябре 2021 года — тогда В. В. Путин потребовал убрать все открытые свалки в черте городов в ближайшие годы<sup>28</sup>. Интерес к спортивным площадкам возрастает с приближением теплого времени года, а к зиме — спадает. Обсуждения высотной застройки и эко-привычек за год уменьшились. Активнее всего строительство высоток волновало горожан в октябре-ноябре 2021 года, а эко-привычки — в сентябре 2021. Пик обсуждений проблем и особенностей безбарьерной среды для маломобильных граждан также приходится на сентябрь 2021 года, однако активно вопросы доступной среды обсуждались и в декабре 2021 года — тогда проходила всероссийская акция Тотальный тест «Доступная среда»<sup>29</sup>; и в июне 2022, когда в силу вступили 5 государственных стандартов, регламентирующих требования, на основании которых осуществляется создание доступной среды<sup>30</sup>; кроме того в июне 2022 Правительство РФ приняло постановление о федеральных грантах для государственных и частных лагерей на создание условий для отдыха детей с инвалидностью и различными нарушениями здоровья.

Динамика обсуждений вопросов, касающихся здоровья города, в социальных медиа за год с сентября 2021 по август 2022 ##### \*Безбарьерная среда

Для более детального анализа редакторской журнала была выбрана тематика безбарьерной среды. Примечательно, что безбарьерная среда воспринимается пользователями не только как физические условия для передвижения маломобильных граждан в городе, но и социальная интеграция инвалидов и других незащищенных групп населения в городскую жизнь через проведение специальных

<sup>28</sup>Путин потребовал убрать все открытые свалки в черте городов: [Электронный ресурс]. URL: <https://iz.ru/1231275/2021-10-05/putin-potreboval-ubrat-vse-otkrytye-svalki-v-cherte-gorodov> (дата обращения 03.09.2022)

<sup>29</sup>3 декабря 2021 года стартует всероссийская акция Тотальный тест «Доступная среда»: [Электронный ресурс]: Минстрой России. URL: <https://minstroyrf.gov.ru/press/3-dekabrya-2021-goda-startuet-vserossiyskaya-aktsiya-totalnyy-test-dostupnaya-sreda/> (дата обращения 03.09.2022)

<sup>30</sup>5 новых ГОСТов по доступной среде: [Электронный ресурс]: ТифлоЦентр Вертикаль. URL: [https://tiflocentre.ru/news/vstuplenie\\_v\\_silu\\_novyh\\_gostov.php](https://tiflocentre.ru/news/vstuplenie_v_silu_novyh_gostov.php) (дата обращения 03.09.2022)

мероприятий, предоставление условий для жизни, учебы, работы и пр.

В центре обсуждений вопросов безбарьерной среды — *инклюзивная среда для детей-инвалидов*, в том числе строительство специализированных детских площадок, учебных заведений, приспособленных под детей с ОВЗ и передвижения маломобильных граждан; а также *условия для передвижения родителей с колясками*. Другая важная тема — *благоустройство городской среды*, предусматривающее установку пандусов, адаптацию тротуаров для передвижения на колясках, с колясками или на велосипедах, капитальный ремонт медицинских учреждений в соответствии с государственной программой «Доступная среда», создание парковочных мест для инвалидов, которые обеспечивают доступность и комфортность посещения торговых центров, парков и других мест; а также оснащение наземного общественного транспорта местами и креплениями для колясок. Отдельно обсуждаются различные проекты по созданию среды для маломобильных граждан, с участием НКО, государственную поддержку инициатив, направленных на создание безбарьерных условий среды.

К социальному компоненту обсуждения безбарьерной среды относятся различные мероприятия для инвалидов, а также научно-просветительские лекции, которые напоминают о присутствии в социуме маломобильных граждан; отдельно обсуждаются налоговые льготы для музеев, театров, библиотек и других культурных учреждений, которые проводят выездные культурные мероприятия для людей с инвалидностью, сирот и пожилых граждан. Важным представляется финансовая поддержка инвалидов, в том числе предоставление им различных льгот и пособий, возможности получать образование (в частности, обсуждается поправка в закон «Об образовании», которая предусматривает получение бесплатного второго профессионального образования соответствующего уровня по иной профессии при наличии инвалидности) и устраиваться на работу.

Структура обсуждений вопросов безбарьерной среды в социальных медиа за год с сентября 2021 по август 2022. Цвета репрезентируют разные блоки обсуждений, выделенные на основе автоматической кластеризации сети. Связь между парой слов — их семантическая близость как минимум в 200 упоминаниях

Figure 41: Структура обсуждений вопросов безбарьерной среды в социальных медиа за год с сентября 2021 по август 2022. Цвета репрезентируют разные блоки обсуждений, выделенные на основе автоматической кластеризации сети. Связь между парой слов — их семантическая близость как минимум в 200 упоминаниях

## 11.2 Обсуждение и выводы

С содержательной точки зрения проведенный анализ показывает, что границы тематики здоровых городов на сегодняшний день задают ряд ключевых подтем. Так, в последние десятилетия в ней активно обсуждались вопросы о важности эмпирического изучения факторов, влияющих на здоровье, социальной справедливости в отношении здравоохранения и решения проблем с доступом к медицине через медицинское обслуживание в городских сообществах. А один из ключевых концептов последнего десятилетия — «риск», который необходимо нивелировать с тем, чтобы создать безопасную городскую среду. При этом эта область является достаточно подвижной и чувствительной к актуальным социальным проблемам и проблемам в области здравоохранения. Именно поэтому на сегодняшний день значимое место в актуальной повестке занимает тема борьбы с эпидемией COVID-19 и последствиями глобальной пандемии. Появление программных работ и обсуждение методологии исследований по тематике здоровых

городов указывает на определенную зрелость этого направления и то, что оно формирует относительно самостоятельную предметную область.

С точки зрения используемой методологии, работа с различными источниками данных (поиск по ключевым словам в WoS; по статьям в конкретном журнале в WoS; сбор упоминаний в социальных медиа) и использование различных методов анализа (анализ ключевых слов, публикаций, тематическое моделирование и пр.) позволили получить более надежные и разнообразные результаты. В частности, выводы на базе анализа сетей цитирования нашли дополнительное подтверждение в анализе ключевых слов, а анализ данных социальных медиа позволил увидеть, как «здоровье» города обсуждается в публичном дискурсе. ## Приложения

Все приложения к настоящей статье, а также подробное описание методологии исследования доступно в репозитории GitHub по ссылке: <https://github.com/Daria-Maltseva/Sociodigger/wiki>

## 12 5.6 Качественный сетевой анализ в эмпирических исследованиях

Сетевые исследования в социальных науках основываются на понятии сети, которое направлено на структурирование социальных взаимодействий. Формальный сетевой анализ известный как анализ социальных сетей (social network analysis, SNA) или количественный сетевой анализ направлен на выявление глубинных структур. Использование качественных методов в сетевых исследованиях становится все более популярным в социальных науках согласно публикациям в Web of Science [199], [201]. И хотя качественные методы всегда присутствовали в сетевых исследованиях на этапах сбора и первичного анализа данных [160], [268], тем не менее позиция качественного сетевого анализа как обособленного методологического подхода является спорной. Одни исследователи убеждены в невозможности обозначения качественного сетевого анализа как независимой методологии [102], другие считают что сетевой анализ в качественной интерпретации возможен [199], [201], [176]. По моему мнению, качественный сетевой анализ может быть обозначен как отдельный методологический подход, выявляющий глубинные смыслы отношений в персональных сетях взаимодействий.

Сегодня качественный сетевой анализ может иметь разные варианты названий, как например эго-сетевой анализ или качественный подход в эго-сетевом анализе, тем не менее, смысл данного методологического подхода кроется именно в анализе глубинных смыслов отношений в сети. Глубинность смыслов отношений понимается как суть и восприятие отношений, не лежащих на поверхности, а проявляющихся в глубине. Качественный сетевой анализ изучает персональные сети или эго-сети. Существует разнородная практика использования качественного сетевого анализа в стратегии смешанных методов [199]. Однако в русскоязычном сегменте нет примера применения только качественного сетевого анализа на практике, чему и посвящена данная работа.

Данный раздел раскрывает возможности и ограничения методологии качественного сетевого анализа в эмпирических исследованиях на примере изучения транзита к родительству в условиях релокации. Данный объект интересен по ряду аспектов: с точки зрения методологии, у релокантов меняются связи и отношения в период адаптации на новом месте, а также в связи с первым опытом родительства взаимодействия также приобретают новый фокус, и с точки зрения актуальной повестки, методология качественного сетевого анализа позволяет обозначить социальные круги отношений, новые смыслы этих отношений и стратегию адаптации к новой жизни в родительстве и релокации. Транзит к

родительству и одновременная адаптация в релокации становится сложным и многогранным процессом для россиян, которые переехали в новые страны после февраля 2022 года. Условия релокации и родительства меняют привычный социальный круг взаимодействий и отношения между индивидами – супругами, родителями, друзьями и.т.д. Качественный сетевой анализ способен раскрыть глубинные смыслы отношений у релокантов и охарактеризовать изменения в социальном кругу до и после рождения ребенка и релокации. Сбор данных происходит при помощи библиографического интервью, в рамках которого собираются и сетевые данные для построения сетевых карт. Также сравниваются два подхода к построению сетевых карт, где в одном случае, сетевые карты строятся исследователем после интервью, а также построение карты предлагается самому информанту. Статья завершается выводами о возможностях и ограничениях качественного сетевого анализа и общими выводами.

### **12.0.1 Транзит к родительству: основные проблемы молодых родителей**

Транзит к родительству («transition to parenthood») обозначает период адаптации, который проходят молодые родители первенца в новой роли. Одни исследователи считают, что этот период начинается после рождения ребенка [44], другие полагают, что адаптация начинается с начала беременности [88], [9]. Транзит к родительству — это переходный период нестабильности и внутреннего конфликта по поводу приобретений и потерь, которые приводят к реорганизации внутренней жизни и поведения [88], [117]. В целом, переход к родительским обязанностям обычно длится с начала беременности до первых нескольких месяцев жизни ребенка, хотя данный период считается нормативным жизненным событием, означающим, что в целом оно ожидаемо и предсказуемо, тем не менее, транзит к родительству накладывает как риски, как на отдельных лиц, так и на всю семью [44], [88]. Белски выделяет четыре типа проблем, с которыми сталкиваются пары во время адаптации к родительству: 1) физическое бремя ухода за младенцем, 2) напряжение в отношениях между мужем и женой, 3) эмоциональные издержки, связанные с сомнениями в компетентности и родительских обязанностях, и 4) личная изоляция [44]. В исследовании потребностей молодых матерей во время беременности и после нее описаны основные запросы на механизмы поддержки во время беременности и после родов, дородовое просвещение, информация о грудном вскармливании, практическом уходе за ребенком и информация о возможных изменениях в отношениях между супружами [101]. Также со стороны молодых матерей заявлены запросы на вовлечение партнеров-мужчин в дородовой и послеродовой периоды и необходимость учиться и обмениваться опытом с другими новоиспеченными родителями [114].

В процессе транзита к родительству может переопределяться гендерный контракт между партнерами. Гендерное соглашение или гендерный контракт рассматривается как основа общественного договора и как часть современной социальной политики [421]. Он описывает негласный договор, но согласованными правами и обязанностями. Концепция гендерного контракта основана на предположении, что «во всех современных обществах существует исторически сложившийся социокультурный консенсус по соответствующим формам гендерного “правильной” формы пола, разделению труда, форме семьи и способу интеграции двух полов в общество через рынок труда и / или семью» ([342] р. 478, [300]). Пересмотр гендерного контракта в процессе транзита к родительству может быть усложнен в условиях эмиграции и релокации. С одной стороны возможно, что форсажорные обстоятельства релокации несколько традиционализируют семью, особенно с ребенком, концентрируя ее вокруг традиционного

гендерного контракта (муж-кормилец, жена-домохозяйка). Но с другой стороны, также возможно формат релокации провоцирует свои вызовы и те женщины, которые прежде до релокации придерживались смешанного контракта работница-будущая мать, в новых условиях возьмут на себя иные задачи - поиск подработки, построение коммуникации вокруг семьи, деятельное освоение социальных возможностей для всех членов семьи и т.д.

Социальное окружение может полностью поменяться в условиях релокации и транзита к родительству. В исследовании социальных сетей у молодых родителей в Германии, на основе опроса было выявлены четыре разных типа социального окружения и поддержки [243]. Первый тип сети, обозначенный как «семейно-удаленная», состоит в основном из друзей и знакомых. Члены социальной сети не будут оказывать социальную поддержку в случае, если это переходит к родительским обязанностям. Второй тип сети, «поляризованный» сеть относительно велика и неоднородна, т. к. количество родственников, друзей и знакомых примерно одинаково, но социальная поддержка молодых родителей возможна. Третий тип «дезинтегрированной» сети состоит из людей, которые плохо интегрированы в общество, которые часто указывают на то, что у них нет или мало человек в сети и которые, таким образом, не способны активизировать большую потенциальную поддержку в своем ближайшем социальном окружении. Четвертый тип сетей «ориентированных на семью» довольно малочисленны и характеризуются высокой долей долгосрочных и тесных связей с членами нуклеарной семьи (братьями и сестрами и родителями). Люди, вовлеченные в такого рода сети, ожидают сильной сетевой поддержки в случае родительства.

Несмотря на изученность темы транзита к родительству, условия релокации привносят дополнительную сложность для адаптации в новой роли. Возможно, сети релокантов переезжая на новое место, становятся похожими на третий тип «дезинтегрированных» сетей без друзей и знакомых в новом для них городе [243]. Дополнительная сложность в интеграции в местное сообщество может базироваться на временной и непостоянной позиции релокантов, которые могут вернуться в Россию или переехать на новое место.

## **12.0.2 Российские релоканты после февраля 2022 года: общий портрет и основные вызовы**

После событий 24 февраля 2022 года многие россияне покинули страну. Хотя точных статистик нет, но по мнению демографов, оценить масштабы эмиграции можно в диапазоне от 550 до 800 тысяч человек [484]. Релокация в отличие от эмиграции носит временный характер, что означает возможное возвращение на родину. Релокация или эмиграция россиян после 24 февраля 2022 года, является дискуссионной темой как в СМИ, так и в академической повестке. Данной теме посвящен целый выпуск журнала «Социодиггер», где описаны мотивы эмиграции, проблемы адаптации, релокация бизнеса и эмигранты в социальных медиа [12]. По данным Forbs, наибольшее число релокантов числится в странах ближнего зарубежья Грузия, Армения, Казахстан, Кыргызстан, а также в странах, где есть возможности длительного пребывания - Турция, Сербия, Израиль, а также страны Евросоюза [478].

Общий портрет релокантов можно описать следующим образом: «это преимущественно молодые люди от 18 до 40 лет, с высшим образованием, активной жизненной и гражданской позицией, занятые квалифицированным, преимущественно интеллектуальным трудом; а следовательно, обладающие высоким уровнем человеческого капитала» [449]. Авторы выделили традиционные проблемы эмигрантов,

которые обсуждались в Телеграм каналах: поиск жилья (долгосрочная и краткосрочная аренда); поиск работы; осуществление банковских операций (обмен валюты, денежные переводы, открытие банковского счета и получение карты); уплата налогов; транспорт и перевозка (расписание транспорта, прохождение границы и таможенного контроля, перевоз багажа, отправка посылок); оформление документов (загранпаспорт, получение прописки, получение ВНЖ); медицинская помощь и обслуживание; обсуждение местных жителей (в контексте их отношения к приехавшим). Также подробно описаны проблемы российских эмигрантов в Армении, которые автор разделяет на объективные и субъективные, где к первой группе относятся ограниченный рынок труда, жилья и нехватка услуг и сервиса, а ко второй группе нежелание российских релокантов адаптироваться местному образу жизни, учить армянский язык, «а напротив, стремление перенести свой образ жизни на новые реалии» [457].

Немаловажным аспектом адаптации релокантов является поддержание привычных социальных взаимодействий и появление новых социальных связей. Социальные взаимодействия между российскими релокантами было охарактеризовано метафорой ризомы из работ Ж. Делеза и Ф. Гваттари, подразумевая горизонтальный характер взаимодействий: «Ризоматическая структура — это неупорядоченная в какую-либо иерархию, множественная и постоянно меняющаяся сеть. Мигранты разбросаны по всему миру, но продолжают посыпать друг другу сигналы и координироваться друг с другом, коммуницировать с друзьями, родственниками и даже с едва знакомыми людьми как за рубежом, так и в России. Ризоматическая структура этой сети дает мощный источник взаимной поддержки и необходимый запас гибкости для поддержания координации между новыми «номадами», которые потеряли свою прежнюю жизнь, но не знают, где начать новую.» [462]. Однако, по моему мнению, релоканты, как и все, все равно выстраивают некоторую иерархию в процессе взаимодействия, демонстрируя, например, разный уровень дохода, адаптации к новому месту и повестке. Исходя из этого можно опустить метафору ризомы, заменив ее на понятие сети, которое может описать социальные взаимодействия без каких-либо условий и ограничений.

Таким образом, сообщество российских релокантов сталкивается с одними и теми же вызовами в разных странах. При этом релоканты испытывают сложности с адаптацией и включением в культурную среду, поскольку точно не знают надолго ли останутся в стране пребывания. Релоканты отличаются молодым возрастом, образованностью, интеллектуальным трудом, активной позицией и высоким человеческим потенциалом. Российские релоканты и эмигранты стремятся поддерживать свой привычный социальный круг взаимодействия, но также нуждаются в поддержке таких же релокантов. Однако в академической повестке пока что не было особого фокуса на семьях с детьми и на тех, кто стал впервые родителями в новой стране в релокации. Данное эмпирическое исследование закроет эту лакуну в научном дискурсе.

### 12.0.3 Дизайн исследования

Изучение транзита к родительству является довольно изученной темой в социологии [44], [88], [117], [101], [114]. Однако актуальность данного исследования состоит в переживании опыта транзита к родительству среди семей российских релокантов, покинувших страну после февраля 2022 года. Поскольку обстоятельства релокации усложняют транзит к родительству, то центральной проблемой становится адаптация к новой жизни и в качестве родителя, и релоканта одновременно. Социальные

взаимодействия молодых родителей в условиях релокации могут сильно измениться, что также влияет на адаптацию. Цель исследования состоит в изучении транзита к родительству в условиях релокации. К задачам исследования относится выявление стратегий адаптации к родительству и к релокации, описание возможного пересмотра гендерного контракта между супругами и изучение изменений в социальных взаимодействиях молодых родителей в условиях релокации. Информантами являются женщины, переживающие свой опыт родительства и релокации. Используется качественный сетевой анализ для сбора и анализа данных при помощи биографического интервью и сетевых карт. Фокус данной статьи направлен на методологические задачи, связанные с применением качественного сетевого анализа, тогда как содержательные выводы будут описаны в другой работе.

Данная работа нацелена на выполнение методологических задач на уровне сбора данных в виде биографического интервью и сетевых карт, что позволит *охарактеризовать возможности и ограничения* качественного сетевого анализа, а также *сравнить способы сбора сетевой карты*, где в одном случае построение сетевой карты предлагается информанту, а в другом случае самостоятельно дорабатывается исследователем на основе проведенного интервью.

#### **12.0.4 Методология качественного сетевого анализа**

Качественный сетевой анализ определяется как методологический подход, выявляющий глубинные смыслы отношений в персональных сетях взаимодействий. Объектом качественного сетевого анализа являются персональные сети или эго-сети, предметом являются глубинные смыслы отношений в сети. Целью качественного сетевого анализа является выявление глубинных смыслов отношений в сети. Рассмотрев разные способы смешивания методов на этапах сбора и анализа, можно заключить, что когда на этапе сбора и анализа данных применяются именно качественные методы, то в таком случае сетевое исследование является качественным [199]. Для осуществления качественного сетевого анализа, на этапе сбора данных могут применяться те методы, которые могут быть качественно проанализированы - интервью, наблюдение и анализ документов, а также структурированы при помощи построения сетевой карты. Таким образом, качественный сетевой анализ состоит из двух этапов: сбор качественных и сетевых данных, анализ качественных текстовых данных, анализ и визуализация сетевых данных.

Немаловажным компонентом применения качественного сетевого анализа является структурализация отношений между акторами при помощи построения сетевой карты. Это техника структурирования социальных отношений, которая может быть выполнена самостоятельно исследователем, либо предлагается информанту во время интервью. Пример использования сетевой карты можно посмотреть в исследовании социального капитала мигрантов Елена Зоммер и Маркус Гампер, которые используют качественные методы исследования, включая полуструктурные интервью и стандартизованный сбор данных сетевых карт [370]. Подробно описано применение сетевых карт, где рассматривается развитие ИТ-бизнеса на начальном этапе и 10 лет спустя (рис. 1). Для заполнения сетевой карты во время интервью информанту выдают лист бумаги с пустыми концентрированными кругами, куда их просят вписать имена или инициалы важных персон. За генератором имен следует набор стандартизованных интерпретаторов имен, которые задавали вопросы об атрибутах изменений в круге и типе связи с каждым изменением. На основе 62 интервью с самозанятыми (или бывшими самозанятыми) мигрантами из бывшего Советского Союза в Германии были выявлены четыре типа

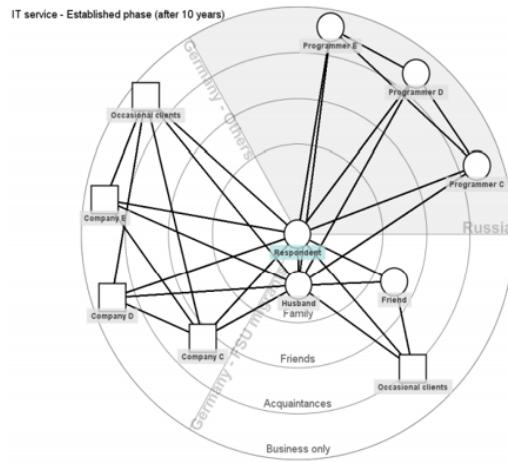
транснациональной предпринимательской деятельности мигрантов [370].

Рисунок 1. Пример сетевых карт бизнес сети ИТ-компании: начальный этап и 10 лет спустя

Figure 1: Business network of an IT company – start-up phase, migrant market



Figure 2: Business network of an IT company – after ten years, mainstream market



### 12.0.5 Примеры транзита к родительству в условиях релокации: кейсы Ольги и Арины

Для демонстрации применения качественного сетевого анализа я предлагаю рассмотреть два кейса семей релокантов, которые переехали в Казахстан. Данные примеры являются периферийными и пограничными с точки зрения адаптации к релокации, где в одном случае демонстрируется сложный адаптационный процесс транзита к родительству, а в другом случае противоположный облегченный опыт. Приводится описание примеров транзита к родительству в условиях релокации, показаны сетевые карты социальных взаимодействий и контекст этих отношений. Оба кейса показывают, как используется качественный сетевой подход в исследовании, выявляют возможности и ограничения данного методологического подхода, а также на основе этих кейсов сравниваются два способа построения сетевых карт – самостоятельное заполнение сетевой карты информантом, либо построение карты исследователем после интервью.

*Кейс Ольги - пример осложненного транзита к родительству в условиях релокации.*

Ольга и Федор жили в Москве, работали в нефтяной отрасли и в ИТ. После начала мобилизации Федор покинул страну из-за рисков попасть на войну, уехав в Кыргызстан, а в это время Ольга оставалась в Москве и готовилась к переезду к мужу. Информанты тщательно и выбирали страну временного проживания, поскольку приближалась дата родов, таким образом, самым важным критерием выбора страны стало качественное медицинское обслуживание, в частности роддом. Спустя полтора месяца Федор и Ольга приехали в Казахстан, нашли жилье, выбрали роддом и подготовились к родам. Дочь родилась на 36 неделе беременности. После рождения дочери у Ольги диагностировали послеродовую депрессию.

Сетевые карты Ольги построены исследователем после интервью. На рисунке 2 изображены две сетевые карты до и после рождения ребенка и релокации. Слева изображена карта до рождения ребенка и релокации, где черным цветом в центре изображено само это у кого брали интервью, в данном случае это Ольга, затем зеленым цветом выделены члены семьи, оранжевым – друзья и голубым – коллеги. Справа представлена сетевая карта Ольги после рождения ребенка и релокации, где фиолетовым цветом добавлен

психолог. Рассказывая о своем опыте родительства и релокации, Ольга отмечала, что именно после рождения ребенка у нее стали меняться отношения с людьми. Например, до рождения ребенка, она на одном уровне воспринимала отношения с родителями мужа, однако после рождения дочери, ее отношения со свекровью ухудшились. Также она близко общалась с подругами Ольгой и Мариной, но после рождения ребенка она стала чуть меньше общаться с Ольгой, но появилась новая подруга Арина. Общение с другими друзьями и друзьями мужа практически прекратилось, теперь Ольга периодически общается с соседками-мамочками из своего дома. При этом Ольга подмечает, что если раньше круг общения состоял из людей такого же уровня образования, возраста и дохода, то теперь взаимодействия носят несколько случайный характер по этим характеристикам, но есть потребность в общении с такими же молодыми мамами как она. Ольга нечасто общалась неформально с коллегами, в связи с чем после выхода в декрет она перестала общаться с коллегами вовсе. После родов возникла потребность в психологической поддержке и с тех пор, она регулярно работает с психологом.



Рисунок 2. Сетевые карты Ольги до и после рождения ребенка и релокации

Характерной чертой осложненного транзита к родительству является послеродовая депрессия. В условиях релокации, возможно, что вероятность данного осложнения вырастает, поскольку женщине нужно время, чтобы адаптироваться к новой среде, подготовиться к родам и комфортному родительству. В случае с Ольгой первое впечатление от Казахстана, города Алматы и их квартиры было очень высоким. Ольга описывает то время, как «эйфорию», оттого что они наконец-то встретились с мужем, нашли жилье и начали готовиться к родам. Сами роды прошли в комфортном роддоме, где она ощущала поддержку врачей и медицинского персонала. Но после родов Ольга стремительно погрузилась в глубокую депрессию: «*Вроде бы счастье такое случилось, но одновременно, появилось ощущение, что вся твоя жизнь рухнула. И ты стоишь на руинах своей жизни, а новую ты еще не построила.*» Ольге очень не хватало поддержки со стороны родителей, близких подруг, но поскольку рядом никого не было, Ольга работала вместе с психологом. Спустя месяц после родов к Ольге приехала на выходные подруга Марина из Москвы. Именно после приезда подруги Марины Ольга ощутила тепло и поддержку, начала выздоравливать, и депрессия начала отпускать. В интервью Ольга подчеркивала, насколько этот кризисный период сблизил ее с подругой Мариной, что теперь их отношения стали крепче чем раньше.

#### ***Кейс Арины – пример облегченного транзита к родительству в условиях релокации***

Арина вместе с мужем переехали в Казахстан из Новосибирска в марте 2022 года, когда Арина

была на 16 неделе беременности. Оба работают в сфере ИТ. Поскольку муж Арины – казахстанец и у него было свое жилье в Казахстане, то они легко решились на переезд. У Арины было много времени на адаптацию на новом месте: вместе с мужем сделали ремонт в своей квартире, получила вид на жительство, прошла курсы для беременных, познакомилась с другими беременными. Также Арина с мужем переезжали в Казахстан в компании друзей. Сын родился на 39 неделе беременности. На выписку из роддома приехали родители и сестра Арины, мама мужа и друзья Арины.

Сетевые карты Арина построила самостоятельно в процессе интервью. На рисунке 3 изображены две сетевые карты до и после рождения ребенка и релокации, где все узлы отмечены одним цветом. Слева изображена карта до рождения ребенка и релокации, а справа представлена сетевая карта Арины после рождения ребенка и релокации. Судя по сетевым картам, узлов и взаимодействий у Арины стало больше после рождения ребенка и релокации. Однако в интервью Арина рассказывала, что общения стало меньше после релокации и рождения ребенка, но появилось больше новых знакомств и общения со старыми знакомыми, которые тоже переехали в Казахстан. До рождения ребенка в кругу самых близких у Арины были муж и две подруги, но после рождения ребенка и релокации с одной подругой они отдалились друг от друга, но зато познакомилась и сблизилась с новой подругой. Довольно интересно, что отношения с мамой, сестрой после рождения ребенка стали более близкими. «*В беременность и после родов я стала больше общаться с мамой и теперь я понимаю, насколько ей возможно было трудно и тяжело со мной и моей сестрой. Благодаря материнству я смогла взглянуть на маму не с позиции ребенка, а с позиции такого же взрослого*». А также с мамой мужа отношения стали более близкими, тогда как раньше они редко общались, поэтому ее нет на сетевой карте слева.

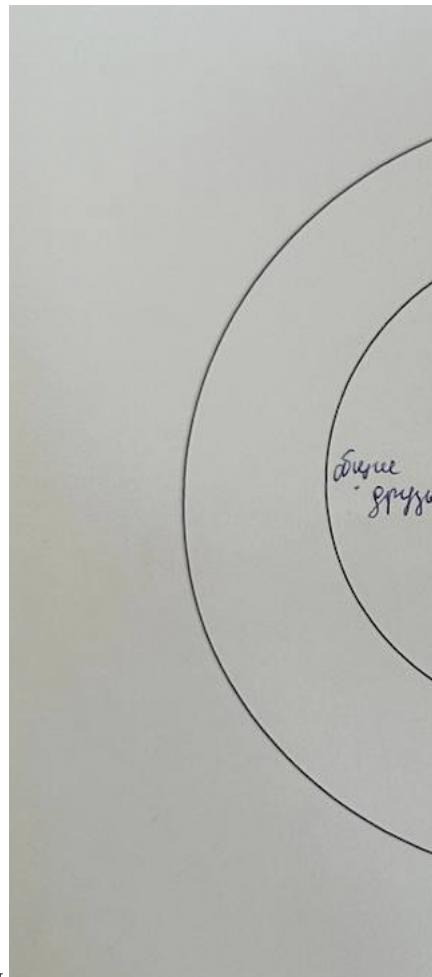


Рисунок 3. Сетевые карты Арины до и после рождения ребенка и релокации

### Рисунок 3. Сетевые карты Арины до и после рождения ребенка и релокации

До переезда Арина с мужем часто общались в компании общих друзей, но после переезда они стали реже контактировать. Но некоторые друзья тоже переехали в Казахстан, и с ними Арина встречается и общается регулярно. Также Арина рано вышла из декрета и работает удаленно несколько часов в неделю. Если раньше она общалась с двумя коллегами, то теперь регулярно взаимодействует только с одной. Также в Казахстан переехали коллега Арины со своей семьей, хотя ранее они часто не общались, то теперь периодически поддерживают связь. Отдельно Арина отметила онлайн-чаты мамочек, в которых общается группа девушек, которые вместе проходили курсы для беременных и другие чаты с мамами в Казахстане. Арина часто обращается в эти чаты за советом и поддержкой, т. к. все их дети ровесники и все сталкиваются с одними и теми же проблемами.

#### 12.0.6 Возможности и ограничения качественного сетевого анализа

Качественный сетевой анализ состоит из двух этапов: сначала интерпретативная часть анализа, а затем ее структурная визуализация [4]. Ранее в систематическом обзоре были рассмотрены возможности качественного сетевого анализа, выделенными в статьях из WoS [199]. На уровне объекта исследования качественный сетевой анализ изучает персональные сети отношений индивидов, а также дает возможность доступа к труднодоступным группам респондентов. На уровне предмета исследования качественный сетевой анализ позволяет изучать глубинные смыслы отношений в сети и контексты взаимодействия, описывать и понимать сети изнутри и снаружи, фокусироваться на деятельности акторов и их стратегиях построения сети, а также выявлять темпоральность отношений в сети. Однако в рамках данной статьи предпринята попытка охарактеризовать возможности и ограничения качественного сетевого анализа на основе конкретного эмпирического исследования. Ниже описана методологическая рефлексия качественного сетевого анализа на основе исследования транзита к родительству в условиях релокации.

Исследование транзита к родительству в семьях российских релокантов выполняется при помощи качественных методов сбора и анализа. Данные собираются методом библиографического интервью, в котором есть структурный блок для построения сетевой карты. Для осуществления качественного сетевого анализа проводится интерпретация данных и визуализация структуры отношений информанта. При желании, собранные сетевые данные можно конвертировать в цифровой вид для последующего количественного эго-сетевого анализа. Тогда исследование будет со смешанным дизайном с конвертированием данных [8].

Сильной стороной качественного сетевого анализа является ориентация на точку зрения индивида при одновременном учете структурного аспекта как части аналитической структуры отношений. Изучая транзит к родительству в условиях релокации, я использовала библиографическое интервью для сбора данных. Дополняя интервью структурным аспектом, индивид рефлексирует не только по поводу пережитого опыта, но и анализирует по поводу случившихся изменений в отношениях с другими индивидами. Например, в интервью с Ольгой, ею было подмечено, что начиная с беременности, она начала меньше общаться с друзьями, друзьями мужа и знакомыми, которых можно отнести к слабым связям. Причем причиной этому стало не только состояние здоровья, но и нежелание навязывать новые для компании темы и продвигать свои интересы.

Качественный сетевой анализ позволяет погрузиться вглубь отношений с каждым узлом,

благодаря чему можно не только визуализировать и структурировать сеть, но и наполнять ее контекстом. На примере отношений Ольги и ее супруга, их отношения можно охарактеризовать как командные и партнерские, поскольку супруги вместе проходили через роды, и теперь, имея удаленную работу, оба ухаживают и заботятся о дочери. Исходя из пересмотра ролей в гендерном контракте, ранее супруги оба работали и приносили доход и бытовые задачи разделяли между собой, то теперь роль «добытчика» легла на плечи мужа, а быт и уход за ребенком на Ольге. Ольга также отметила перемены в муже после рождения ребенка: *«Он сполна ощутил ответственность за финансовое положение нашей семьи в первые два-три месяца после рождения ребенка. Тогда курс рубля к тенге стал хуже, наш доход упал. Из-за чего Федя стал все свое время и внимание инвестировать в поиск новой работы – изучал английский, откликался на вакансии, проходил интервью. Спустя несколько месяцев он нашел новую работу с большим окладом и успокоился»*. Хоть на сетевой карте и не видна разница в отношениях Ольги и Федора, но исходя из интервью их отношения стали крепче и перешли на новый более сложный уровень.

Другая явная особенность качественного сетевого анализа заключается в возможности изучать отношения с точки зрения темпоральности. В данном случае информанты описывали свои отношения до и после рождения ребенка и релокации. И хотя не обозначено конкретных временных промежутков и точных дат, наглядно видно, как эти два события поменяли социальный круг взаимодействий. По условной шкале транзита к родительству и адаптации в релокации описанные примеры являются полярными по отношению к друг другу, где в одном случае показана низкая адаптивность, а в другом высокая. При этом в каждом из этих пограничных случаях наблюдается разительное изменение в социальном круге после рождения ребенка и релокации. В случае с Ариной, у которой была сравнительно легкая адаптация в релокации, то в ее социальном окружении увеличилось количество узлов после переезда и рождения ребенка. Возможно, это искажение случилось впоследствии релокации, т. к. Арина засчитала не все связи и взаимодействия, которые были до переезда. Вероятно, у Арины выделен несколько смещенный фокус на знакомых людей в Казахстане, связь с которыми является значимой в условиях эмиграции.

К ограничениям качественного сетевого анализа относится общую субъектность сети, которая строится на основе восприятия своих контактов и связей. Информант может не всех вспомнить и не отметить всех на сетевой карте. Например, в сетевой карте Арины случилось искажение в увеличении количества узлов после релокации, что можно отнести к ограничениям качественного сетевого анализа. Другим ограничением методологии может быть возможное расхождение сетевой карты и интервью, если информант самостоятельно заполняет карту. При этом могут быть проблемы в понимании информантом как правильно нужно заполнять сетевую карту. Например, в случае с Ариной, она не указывала имен в сетевой карте, хотя в интервью переходила на имена. Однако после окончания интервью попросила не указывать настоящих имен и изменить ее имя в том числе.

Ограничением эмпирического исследования является сложный рекрут информантов. Приглашения на участие в исследовании распространяются через телеграм-чаты релокантов, но отклик происходит сравнительно медленно. Также для полной картины для изучения транзита к родительству, правильно было бы опрашивать новоиспеченных отцов в том числе. Однако, пока что отклик со стороны мужчин собрать сложно, но, несомненно, это ограничение существенно.

## 12.0.7 Сравнение способов сбора сетевых карт

Сбор сетевых карт возможен двумя способами: самостоятельное заполнение информантом или сетевая карта строится самим исследователем, где он полагается на данные из интервью, наблюдения или документов [176]. В рамках данной статьи ставилась задача сопоставить эти подходы к сбору сетевых карт и выбрать более подходящий способ для изучения транзита к родительству.

Построение своей сетевой карты информантом – это рефлексивное и вдумчивое упражнение. Поскольку интервью проводится онлайн по зуму, то нужно выделить время на построение сетевой карты и после на обсуждение этой карты. Безусловно, сам процесс построения сетевой карты интересный и затягивает самого информанта, однако не всегда хватает времени на полноценное обсуждение. Также тратится время на постановку задачи построения сетевой карты и может быть сложно донести образ конечного результата. Помимо этого, есть предположение что сбор сетевой карты онлайн и по зуму может быть сложнее, чем вживую. Как итог, сетевую карту информант отправляет исследователю и качество картинки может быть не таким высоким или на карте присутствуют помарки, подчерк может быть непонятным и др. Однако, построение сетевой карты информантом как упражнение может быть полезным, т. к. интервью может не охватить каких-то людей или не обозначить изменения в отношениях с другими узлами. Получается, что благодаря самостоятельному построению сетевой карты информантом сглаживаются ограничения сбора данных посредством интервью.

Другим способом построения сетевой карты является заполнение сетевой карты самим исследователем после интервью. Этот способ сбора сетевых данных является более экономичным, не тратится время информанта, а также исследователь визуализирует сетевую карту как нужно для его исследования. Для построения сетевой карты необходимо лишь задать дополнительные вопросы о социальном окружении информанта. Визуализация полностью зависит от исследователя, поэтому, вероятно, она будет высокого качества и наглядно структурировать все контакты. Однако в данном случае, исследователь в большей степени полагается на интервью, в ходе которого информант может вспомнить не всех узлов или отметить изменения не во всех отношениях.

В случае данного эмпирического исследования способ самостоятельного заполнения сетевой карты информантом показал большую результативность. И хотя на данный способ сбора сетевых данных тратится больше времени информанта, но зато можно выявить ранее незамеченные свойства социального окружения и его возможные искажения. Как например, в случае с Ариной, где после релокации в сетевой карте появляется большее количество узлов. Таким образом, данный способ сбора сетевых данных является приоритетным для изучения транзита к родительству в условиях релокации.

Таким образом, качественный сетевой анализ является сравнительно молодой методологией, по данным WoS рост публикаций с использованием качественных методов в сетевых исследованиях начался с 2010 годов [199], [201]. Однако в русскоязычном поле еще не было работы с практическим применением качественного сетевого анализа. В данном разделе приведены два полярных друг к другу кейса с точки зрения адаптивности к условиям релокации. В одном случае описан усложненный кейс транзита к родительству через релокацию накануне родов, что привело к послеродовой депрессии и более длительной адаптации в новой роли и условиях. В другом случае показана упрощенная адаптация в новой стране, с ранним переездом, наличием собственного жилья, подготовкой к родам и более широкому кругу поддержки. На основе двух этих примеров раскрываются возможности и ограничения методологии

качественного сетевого анализа, а также сравниваются подходы к построению сетевых карт.

К возможностям качественного сетевого анализа относятся глубинное погружение в отношения в сети информанта, насыщение контекстом этих отношений и их темпоральный анализ. Поскольку в условиях релокации круг социального взаимодействия меняется, то транзит к родительству может быть усложнен. Среди ограничений качественного сетевого анализа можно отнести субъектность построенной сети, которая может отличаться ошибками иискажениями. Сама сетевая карта без погружения в контекст не до конца раскрывает специфику отношений.

Сравнивая два подхода к построению сетевых карт, можно заключить, что самостоятельное заполнение информантом может принести больше рефлексии по поводу отношений, однако сама процедура является затратной по времени и ресурсам. При этом есть недочеты в конечной визуализации – оформление, помарки, непонятный подчерк, фотография низкого качества и др. Другой подход к построению сетевой карты основывается на сетевых данных, собранных из интервью. К преимуществам данного подхода можно отнести качество визуализации, в которой будет наглядно структурированы все контакты в том виде, в каком исследователь захочет, а также в данном случае не тратится лишнее время информанта.

## 13 Применение ERGM для анализа конференций

### 13.1 Введение

В последние годы концепция “умного города” привлекает все большее внимание, поскольку городские центры сталкиваются с проблемами, связанными с быстрой урбанизацией, нехваткой ресурсов и необходимостью устойчивого развития [174, 207]. В умном городе используются передовые технологии и подходы, основанные на данных, для оптимизации городской инфраструктуры, повышения качества жизни горожан и обеспечения устойчивого экономического роста. По сути, эта концепция воплощает в себе видение технологически обусловленной городской утопии, направленной на преобразование и улучшение городской среды [396]. Хотя глобальное понимание умного города в первую очередь связано с внедрением интеллектуальных технологий и аспектов искусственного интеллекта, важно признать, что этот термин включает в себя не только цифровизацию. Т. Нам и Т. Прадо [283] придерживаются комплексного подхода к этому понятию, охватывающего три аспекта: технологии, людей и сообщества/институты. Таким образом, каждый город ставит большие задачи перед людьми и институтами, ответственными за его прямое или косвенное управление, планирование и модификацию [323].

В силу своей междисциплинарной природы концепция охватывает различные области - городское планирование, внедрение инноваций, устойчивое развитие и менеджмент. Академические дискуссии вокруг темы “умного города” по своей природе характеризуются междисциплинарностью. Среди упоминаний “умного города” доминируют области ИКТ, экологии и энергетики, а также урбанизации [186]. Сложность этих дискуссий обусловлена разнообразным и многообразным использованием термина, а также его связями со смежными понятиями (например, интеллектуальный, цифровой, информационный) [174]. Как отмечает А. Коккиа, “концепция умного города до сих пор не имеет единого определения и по-разному интерпретируется в зависимости от области внимания” [79].

Помимо присущей этому термину неоднозначности, практическое понимание умного города и целей, лежащих в основе программы применения концепции в России, также имеет свои особенности. Крупные российские агломерации активно участвуют в разработке инициатив, охватывающих различные сферы, таких как: здравоохранение, образование, утилизацию отходов, сбор данных и мониторинг. Москва, Санкт-Петербург и Казань достигли стадии “Умный город 3.0”, характеризующейся участием граждан в инновациях [403]. Россия представляет собой крайне особенный пример реализации концепции в силу значительной концентрации городского населения и относительно небольшой доли населения, проживающего в сельской местности. Неосведомленность граждан является существенным препятствием для развития “умных городов” в России [431].

Однако настоящее исследование отвечает на другой вопрос, не связанный с концептуализацией термина. Его цель - проанализировать, какие институты вносят вклад в развитие концепции “умного города” в российском академическом контексте. В исследовании рассматривается процесс сотрудничества в российском научном сообществе, выявляются его закономерности, ключевые участники и анализируется структура такого взаимодействия. В отличие от наукометрического подхода к рассмотрению термина “умный город”, использующего библиометрические методы для анализа сетей цитирования [186, 323, 378], данное исследование сосредоточено на анализе компаний - участников научных мероприятий, проходивших в России с 2007 по 2022 год. Выборка была получена смешанным методом, как парсингом, так и ручным сбором данных с платформы *eLIBRARY.RU* - крупнейшей российской полнотекстовой базы данных научных журналов от ведущих академических, университетских, отраслевых и коммерческих издательств.

Мы работали с тремя исследовательскими вопросами:

1. Кого можно назвать ключевыми игроками, способствующими развитию концепции “умного города” в российском научном сообществе?
2. Какие сообщества можно выделить в сети сотрудничества между организациями?
3. Каковы ключевые характеристики сети сотрудничества среди организаций, участвующих в формировании повестки умного города?

## 13.2 Данные и методы

С платформы *eLIBRARY.RU* использовались документы типа “материалы конференций” с поисковым запросом “умный город”. Нас интересовали все материалы, в которых данная тема упоминалась в названии статьи или в ключевых словах публикации. Общее количество 652 научных мероприятий было собрано вручную, поскольку единого подхода к сбору такого рода данных из первоисточника не существует. Список уникальных участников лег в основу второй группы вершин сетевых данных - списка компаний. Всего было получена 2761 уникальная организация. После процедуры чистки данных количество было уменьшено до 2211 уникальных значений.

Итоговый набор данных состоит из мероприятий ( $n = 652$ ) и компаний ( $n = 2211$ ), которые принимали участие в этих событиях. Для формирования сети взаимодействия был создан список ребер (*edgelist*) (первый столбец назван *from*, второй - *to*). Под бимодальными данными понимаются данные, фиксирующие связи между двумя разобщенными группами, поэтому связи между этими группами

отсутствуют. Кроме того, для обоих групп вершин были созданы файлы атрибутов. С помощью программы *txt2rajek 3* (Pfeffer et al., 2014) файл со списком ребер размером  $6186 \times 2$  был преобразован в бимодальную сеть ***EC***, состоящую из событий  $\times$  компаний. Затем бимодальная сеть была разбита на две одномодальные сети. Сеть “Компании” (Companies) была получена умножением исходной сети ***EC*** на транспонированную ***CE***, а сеть “Мероприятия” (Events) - умножением ***CE*** на ***EC***. В итоге для дальнейшего анализа мы сформировали 3 отдельные сети и присвоили их вершинам соответствующие атрибуты (табл. 1). Множественные петли и петли (loops) были удалены.

	# Nodes (sum)	# Mode 1	# Mode 2	# Edges
EC	2,863	652	2,211	6,186
Events	652			22,573
Companies	2,211			52,824

Для изучения двудольных сетей традиционно применяются несколько подходов. Наиболее распространенный заключается в преобразовании бимодальных данных в одномодальные, что позволяет применять традиционные методы сетевого анализа [51]. Такое преобразование достигается путем создания сети, в которой узлы представляют вершины из одной группы, а ребра - связи между вершинами из другой группы (например, компаниями и мероприятиями). Исследование сплоченных групп на одномодальных сетях позволит выявить основные группы участия в сети коллaborаций. Под сплоченными подгруппами понимаются кластеры или подмножества узлов в сети, которые демонстрируют более высокую степень взаимосвязанности между собой по сравнению с остальной частью сети [414]. Эти подгруппы характеризуются сильными внутренними связями и относительно более слабыми связями с узлами за пределами подгруппы. Рассматривая сплоченные подгруппы в сети взаимодействия компаний, мы остановимся на следующих: клики (cliques) [414], к-ядра [309], Лувенский алгоритм [50] и Лейденский алгоритм [398].

Несмотря на то что единый подход к анализу бимодальных данных в анализе социальных сетей основан на их преобразовании в одномодальные, в последние десятилетия исследователями была разработана отдельная группа методов, позволяющих анализировать обе группы одновременно [51, 52]. Для непосредственного анализа бимодальных данных будут использоваться “хабы и авторитеты” (hubs and authorities) Кляйнберга [209], бимодальные ядра (two-mode cores). Кроме того, для ответа на первый исследовательский вопрос будут использованы три меры центральности (degree centrality, betweenness centrality и eigenvector centrality).

Для ответа на третий исследовательский вопрос мы используем экспоненциальное моделирование случайных графов (Exponential Random Graph Modelling, ERGM) и сетевой автокорреляции для исследования структур зависимости в сетях сотрудничества. ERGM представляют собой подход к статистическому моделированию, который позволяет исследователям изучать закономерности связей или отношений в этих сетях. Применяя ERGM, мы выявим основные механизмы и процессы, управляющие формированием связей, и изучим факторы, определяющие формирование отношений сотрудничества.

## 13.3 Результаты

### 13.3.1 Исследовательский вопрос 1

Первоначально использовался бимодальный вариант узлов и авторитетов Клейнберга [209]. Используя программу Pajek, мы задали количество вершин из первой и второй группы равным 10. С помощью этой операции мы создаем подсеть. Здесь интересны несколько моментов. Во-первых, большинство мероприятий являются международными научными конференциями, поскольку они потенциально привлекательны для политиков, исследователей и профессионалов отрасли. Во-вторых, все организации являются ведущими государственными университетами. Эти университеты можно рассматривать как ключевые движущие силы обмена знаниями.

Другим распространенным подходом при анализе бимодальной сети напрямую является применение бимодальных ядер [63]. Мы выбрали пороговые значения  $p = 10$  и  $q = 6$ . Эти значения образуют разбиение, которое дает возможность выделить подсеть журналов, в которых не менее 10 компаний (в этой подсети) принимали участие в каком-либо мероприятии, и мероприятия, в которых участвуют не менее 10 компаний (в этой подсети). Большинство мероприятий - это международные конференции, которые проводились в Москве. Большинство компаний были крупными (размер  $> 3$ ). Из этого можно сделать вывод, что как для участия в мероприятиях, так и для их организации необходимы весомые источники, и только крупные компании могут привлекать участников из влиятельных и центральных с точки зрения положения в сети компаний.

Далее к бимодальной сети были применены меры центральности. Под степенной мерой центральности общего количества связей мы будем понимать количество связей, которое имеет вершина. Она позволяет количественно оценить уровень вовлеченности и активности акторов в сети. Более высокое значение указывает на большее количество связей или более высокую вовлеченность компании в различные события. Анализируя сеть, можно отметить центральное положение российской компании “Ростелеком” (Comp1721) в сети и ее значительную вовлеченность в многочисленные события (degree centrality = 67). Высокая степень центральности говорит о том, что эта компания установила прочные связи, партнерские отношения и аффилированность с широким кругом мероприятий в области. Другая важная компания - Российская академия народного хозяйства и государственной службы при Президенте РФ (РАНХиГС) (Comp1648) - также может рассматриваться как ключевой игрок, имеющий большее число связей или вовлеченний компаний в различные мероприятия (degree centrality = 65). Эти компании можно рассматривать как локальные центры в промышленности и науке.

Центральность по посредничеству организации НИУ ВШЭ (Comp1383) характеризует ее как компанию с высоким потенциалом контроля над потоком информации между другими компаниями и конференциями (центральность по посредничеству = 572569,3). Показатель определяет степень, в которой компания выступает в качестве связующего звена или посредника, способствующего коммуникации и сотрудничеству в сети (Wasserman & Faust, 1994). Другим важным участником сети с точки зрения центральности по посредничеству является компания “Ростелеком”, которая играет важнейшую роль в обеспечении связи между другими компаниями и событиями (центральность по посредничеству = 441545,6). Российская академия народного хозяйства и государственной службы при Президенте РФ (РАНХиГС) (Comp1648), опять же, может рассматриваться как ключевой игрок (центральность по посредничеству = 348313,1), а также Университет ИТМО (Comp990) (центральность по посредничеству

= 236246,4).

С точки зрения центральности по собственному вектору наибольшее значение имеет Российская академия народного хозяйства и государственной службы при Президенте РФ (РАНХиГС) (Comp1648), что свидетельствует о ее высокой значимости и влиянии в домене “умный город” (центральность по собственному вектору = 1,0). Аналогичным образом в качестве важных вершин рассматриваются Российский экономический университет имени Г.В. Плеханова (Comp1747) (центральности по собственному вектору = 0,96) и Финансовый университет при Правительстве РФ (Comp2116) (центральность по собственному вектору = 0,91). Эти компании связаны со значительным числом важных и влиятельных событий, что свидетельствует об их ключевой роли в содействии сотрудничеству, распространению знаний и инноваций в сети.

Таком образом, ключевыми игроками, судя по высоким показателям центральности, являются ОАО “Ростелеком”, НИУ ВШЭ, Президентская академия (РАНХиГС), Университет ИТМО, Российский экономический университет имени Г.В. Плеханова и Финансовый университет при Правительстве РФ. Эти организации можно рассматривать как ключевые движущие силы обмена знаниями, поскольку они занимают выгодное положение в сети коллоквий.

### 13.3.2 Исследовательский вопрос 2

Сеть компаний имеет 2211 вершин, 52824 ребра. Сеть является взвешенной, так как представляет собой проекцию базовой бимодальной сети **EC**. Плотность сети равна 0,022, что означает наличие 2,2% всех возможных связей. Сеть имеет 21 (слабую) компоненту, размер самой большой компоненты равен 2170. Транзитивность (или коэффициент кластеризации) равна 0,354. Веса варьируются от 1 до 24. Диаметр (длина кратчайшего пути) равен 7.

Основная стратегия работы с такими проекциями заключается в том, чтобы выбрать пороговое значение веса ребер и выделить подсеть [51]. Мы решили удалить все связи со значением меньше 7 и удалить все изолированные вершины. Редуцированная сеть содержит 69 вершин и 159 ребер. Плотность редуцированной сети равна 0,068, что означает наличие 6,8% всех возможных связей. По данным рис. 1 можно сказать, что визуализация редуцированной сети имеет некоторые интересные особенности (размер узла зависит от размера компании). Большинство компаний являются государственными, то есть они составляют ядро всей сети. Также примечательно, что сеть имеет разделение: одна часть состоит из университетов, а другой - из коммерческих компаний. Такое разделение позволяет говорить об очевидной дифференциации и разделении типов организаций в сфере “умного города”: университеты занимаются преимущественно исследованиями и академической деятельностью, а бизнес-компании - коммерческой деятельностью и практическим применением.

Визуализация редуцированной сети компаний с атрибутами

Figure 42: Визуализация редуцированной сети компаний с атрибутами

Транзитивность сети равна 0,355. Для оценки значения этого коэффициента был проведен CUG тест (Conditional Uniform Graph) (рис. 2). Учитывая структуру сети (обусловленную количеством ребер и диад), значение 3,355 выше по сравнению со случайными графами той же топологии, что свидетельствует о тенденции к кластеризации компаний. Таким образом, можно утверждать, что наша редуцированная сеть обладает высокой транзитивностью, имея в разы больше треугольников, чем

симулированные сети ( $n=1000$ ). Дальнейшее изучение характеристик, описывающих структуру сети, может способствовать более глубокому пониманию закономерностей взаимодействия, информационных потоков и обмена знаниями между различными заинтересованными сторонами в сфере умного города. Поэтому применение более современных методов представляется чрезвычайно актуальным.

#### Результаты CUG теста

Figure 43: Результаты CUG теста

Наибольшая клика редуцированной сети состоит из 7 вершин (Кубанский государственный аграрный университет, Донской государственный технический университет, Финансовый университет при Правительстве РФ, Сибирский федеральный университет, Северо-Кавказский федеральный университет, Российский экономический университет имени Г.В. Плеханова, Президентская академия (РАНХиГС)). Несмотря на географическую удаленность друг от друга, наличие сплоченной и тесно взаимосвязанной подгруппы среди этих университетов свидетельствует о тесном взаимодействии и общности интересов. Их коллективное участие в сети свидетельствует о совпадении целей, направленности исследований и научной деятельности. Сплоченность этой группы свидетельствует о наличии устоявшейся сети научного взаимодействия и совместных инициатив, выходящих за пределы географических границ.

Затем мы провели анализ для выявления k-ядер в сети (рис. 3). Примечательно, что самыми крупными k-ядрами оказались 3-ядро (зеленый цвет) и 6-ядро (оранжевый цвет). В 3-ядре входят в основном государственные компании, что свидетельствует об активном участии государственных структур в совместной академической деятельности. Данное наблюдение подчеркивает стремление государства к развитию синергетических отношений и облегчению процесса сотрудничества в экосистеме умного города. Напротив, в 6-ядрах и 4-ядрах группах (желтые цвета) преобладают университеты, что свидетельствует об общей заинтересованности и вовлеченности этих учебных заведений в области. В 5-ядре (голубой цвет) представлены университеты с высшим рейтингом, что свидетельствует об их известности и признанной компетентности в данной области. Присутствие этих уважаемых учебных заведений в пятом ядре означает их лидерство и влияние на формирование дискурса, программы исследований и совместных инициатив в предметной области.

#### Визуализация редуцированной сети на основе 3-ядер

Figure 44: Визуализация редуцированной сети на основе 3-ядер

Далее мы применили алгоритмы выявления сообществ (Leiden и Louvain) для анализа подгрупп в сети (рис. 4). Применение алгоритма Louvain позволило разделить сеть на четыре различных сообщества. Светло-голубой кластер состоит в основном из университетов, что подчеркивает их коллективное участие и совместное взаимодействие в изучении концепции “умного города”. Зеленый кластер включает в себя как университеты, так и коммерческие компании, что подчеркивает взаимосвязь и потенциальный обмен знаниями между этими двумя структурами. В синем кластере преобладают государственные компании, что свидетельствует об их значительной роли и участии в совместной деятельности в контексте обсуждения умного города. Наконец, в розовом кластере представлены ведущие российские технические университеты, что подчеркивает их признанный опыт и влияние на формирование дискурса. Рассчитанная модульность равна 0,44. Это значение указывает на относительно высокую степень сплоченности выявленных сообществ, что говорит о том, что алгоритмы обнаружения

сообществ эффективно отражают значимые и отчетливые модели взаимодействия в сети коллaborций. Аналогичным образом алгоритм Лувена дал сопоставимые результаты, но с одним существенным отличием. Он объединил зеленый и синий кластеры, что свидетельствует о более высоком уровне взаимосвязи и совместного взаимодействия между университетами и государственными компаниями. Такое объединение кластеров отражает потенциально более сильную синергию и общие цели этих организаций.

Визуализация редуцированной сети на основе алгоритмов кластеризации Louvain (слева) и Leiden (справа)

Figure 45: Визуализация редуцированной сети на основе алгоритмов кластеризации Louvain (слева) и Leiden (справа)

### 13.3.3 Исследовательский вопрос 3

Для изучения зависимости между связями в сети взаимодействия компаний и размером компаний мы провели сетевой автокорреляционный анализ. В ходе анализа были рассчитаны две широко используемые статистики автокорреляции: Индекс I Морана и Индекс С Гири. Расчетное значение I Морана составило 0,54, что указывает на умеренную положительную автокорреляцию между связями в сети и размером компаний. Это говорит о том, что компании со схожими размерами склонны к сотрудничеству друг с другом, образуя кластеры или группы в сети. Положительная автокорреляция говорит о том, что крупные компании чаще сотрудничают с другими крупными компаниями, а мелкие компании - с мелкими. Аналогичным образом был рассчитан индекс С Гири, значение которого составило 0,22, что также указывает на умеренную положительную автокорреляцию. Это еще раз подтверждает идею о том, что компании схожего размера склонны к сотрудничеству в сети.

Полученные результаты свидетельствуют о том, что размер компании играет определенную роль в формировании моделей сотрудничества в сети. Наблюдаемая умеренная положительная автокорреляция означает, что компании сопоставимого размера с большей вероятностью будут создавать партнерства и участвовать в совместной деятельности, что может быть обусловлено наличием общих ресурсов, опыта или взаимных интересов.

Экспоненциальные модели случайных графов (ERGM) - это статистические модели, используемые для анализа сетевых структур и позволяющие исследователям делать выводы о закономерностях сетевых связей. Результат применения ERGM представлен в табл. 2. Самая простая модель (Модель 1) не представляет особого интереса - она не дает нам ничего сверх простой вероятности образования ребра в существующей сети (вероятность = 0,067). С помощью этой компоненты мы ответим на вопрос: имеет ли сеть тенденцию к образованию взаимных связей? Эта мера равна плотности сети. Положительное значение коэффициента означает, что мы имеем больше ребер, чем в случайной модели; отрицательное - меньше, чем в случайной модели.

Таблица 2. Сводная статистика ERGM для сети взаимодействия

Figure 46: Таблица 2. Сводная статистика ERGM для сети взаимодействия

Далее мы добавляем параметры, основанные на атрибутах, генерируя случайные сети, которые по ним соответствуют наблюдаемой сети. Количество связей внутри компаний схожего размера -

*nodematch('size')* - и схожих значений рейтинга - *nodematch('size')*, - согласно модели 2, говорит о том, что в данной сети нет статистически значимой вероятности этих параметров. *Nodematch* позволяет определить, склонны ли участники связываться с компаниями, имеющими схожий размер. Другими словами, гомофилия по этим признакам в сети отсутствует.

Далее мы включили *nodefactor('gov')* - количество раз, которое узлы с заданным уровнем категориального узлового атрибута встречаются в наборе ребер. Также мы рассматриваем атрибутивную переменную - абсолютное значение разницы в размере компаний - *absdiff('size')*. Следует помнить, что коэффициенты, их стандартные ошибки и р-значения находятся в логарифмической шкале, поэтому мы интерпретируем их как коэффициенты logit.

Результаты *nodefactor* свидетельствуют о том, что в этой сети существует статистически значимая вероятность связи с компаниями, которые являются или не являются государственными. В нашем случае государственные компании связаны в 1,261 раза чаще по сравнению со случайной сетью. Аналогично, компании с размером 3 и компании с размером 4 связаны в 0,55 и 1,27 раза больше, чем случайно, соответственно.

*Absdiff* исследует, образуют ли два узла связь, потому что они оба имеют схожие значения (т.е. меньшую разницу) или потому что они не имеют схожих значений (т.е. большую разницу). Вероятность того, что узлы с одинаковыми оценками будут общаться, в 0,999 раза выше, чем вероятность случайности. Аналогично, вероятность того, что узлы с одинаковым размером будут общаться, в 1,106 раза выше, чем вероятность. Коэффициенты относительно невелики, однако они демонстрируют высокий уровень статистической значимости.

В заключительной части анализа мы оцениваем, насколько модель “подходит к данным”, т.е. насколько хорошо она воспроизводит некоторые глобальные свойства (общая статистика Goodness-of-Fit). Для этого мы выбираем базовую сетевую статистику, отсутствующую в модели, и сравниваем значение этой статистики, наблюдаемое в исходной сети, с распределением значений, полученных в симулированных сетях с помощью нашей модели. Можно с уверенностью сказать, что 3-я модель хорошо согласуется с данными. Выборочная статистика представлена на рисунке 5.

Выборочная статистика итоговой модели

Figure 47: Выборочная статистика итоговой модели

## 13.4 Заключение

Концепция “умного города” находит все большее применение в качестве ответа на различные городские проблемы, обусловленные такими факторами, как рост численности населения, ускоренное экономическое развитие и экологические вызовы. Умный город направлен на решение таких проблем, как преступность, пробки, недостаточное качество услуг и экономические ограничения, предлагая при этом перспективы инклюзивного процветания и повышения благосостояния.

Ключевыми игроками, судя по высоким показателям центральности, являются ОАО “Ростелеком”, НИУ ВШЭ, Президентская академия (РАНХиГС), Университет ИТМО, Российский экономический университет имени Г.В. Плеханова и Финансовый университет при Правительстве РФ. Эти институты можно рассматривать как ключевые движущие силы обмена знаниями, поскольку они обладают связями

с влиятельными и хорошо известными событиями.

Редуцированная сеть обладает высокой транзитивностью. В качестве целостной и тесно взаимосвязанной подгруппы можно рассматривать 7 учебных заведений: Кубанский государственный аграрный университет, Донской государственный технический университет, Финансовый университет при Правительстве РФ, Сибирский федеральный университет, Северо-Кавказский федеральный университет, Российский экономический университет имени Г.В. Плеханова, Президентская академия (РАНХиГС). Результаты выявления сообществ показали, что в сети сотрудничества существует несколько отдельных сообществ (университеты, университеты и частные предприятия, государственные компании, технические университеты высшего уровня).

В процессе анализа удалось установить, что размер компаний играет важную роль в формировании моделей сотрудничества в сети. Наблюдаемая умеренная положительная автокорреляция означает, что компании сопоставимого размера с большей вероятностью образуют партнерства и участвуют в совместной деятельности, что может быть обусловлено наличием общих ресурсов, опыта или взаимных интересов. Результаты моделирования с применением ERGM показали, что государственные компании связаны между собой в 1,261 раза чаще, чем случайно. Аналогично, средние и крупные компании связаны между собой в 0,55 и 1,27 раза больше, чем случайно, соответственно. Вероятность связи между университетами с одинаковым рейтингом в 0,999 раза выше случайности.

## **13.5 5.9 SNA для анализа пользовательского контента: взгляд через призму маркетинга**

### **13.5.1 Рынок электронной коммерции: пользователи и отзывы**

С каждым годом растёт доступность получения информации населениями разных стран. Так, в России по данным на 2021 год более 85 % населения страны имеют доступ к интернету [11]. Одновременно с этим происходят и более глобальные тенденции, например развитие эры Web 4.0, которая является отражением развития интернет-технологий.

Так, всё большее количество процессов и взаимодействий переходят в онлайн среду, а экономика разных стран в свою очередь развивается под влиянием новых цифровых технологий, машинного обучения и искусственного интеллекта, инноваций и др. Электронная коммерция, являясь одним из секторов экономики России, тоже продолжает активно расти, достигнув в 2020 году 2,5 трлн. руб. [473] Пользователи же в свою очередь также начинают проводить всё большее количество времени в онлайн, совершая там разные действия, включая различные экономические транзакции (напр., онлайн-покупки) и поддерживая коммуникацию, как между пользователями, так и с брендами.

Таким образом, в эру онлайн-торговли пользователи совершают не только покупки онлайн, но и после их завершения оставляют отзывы о предоставленных услугах и купленных товарах. Например, онлайн-отзывы в академической литературе зачастую относят к таким понятиям, как пользовательский контент (UGC или user-generated content) и электронное сарафанное радио (e-WOM). В академической и бизнес литературе по-прежнему нет единого устоявшегося определения, поэтому в рамках данной работы под пользовательским контентом мы будем понимать медиаконтент, созданный или произведенный широкой публикой, а не оплачиваемыми профессионалами, и в основном распространяемый в Интернете [94]. Второй термин, электронное сарафанное радио будем приравнивать к онлайн-отзывам, полученным

от онлайн-пользователей [163].

Существует большое количество работ, написанных о пользе электронного сарафанного радио и в целом пользовательского контента, поскольку они зачастую рассматриваются как фактор, влияющий на принятие решений о покупке. Помимо прочего, отзывы и комментарии, оставляемые на онлайн-платформах, рассматриваются через призму такого психологического явления, как социальное доказательство. Поскольку процесс принятия решения потребителем о покупке онлайн становится всё более комплексным, влияние мнение других людей может быть очень сильным при принятии решения о покупке. Потребители поэтому часто опираются на отзывы других пользователей онлайн-ресурсов, рейтинги товара, чтобы оценить качество, функциональность и ценность продукта. На основе проведённого анализа теоретических источников был выявлен перечень основных факторов электронного сарафанного радио, которые могут повлиять на принятие решений в онлайн. К ним относятся:

- Качество отзыва [76]:
  - Достоверность источника [417]
  - Уровень субъективности, читабельность, орфографические ошибки [142]
- Полезность отзыва и воспринимаемая полезность отзыва [204]:
  - Репутация авторов
  - Полнота (насыщенность) отзыва, простота понимания, краткое, сжатое резюме [67]
  - Воспринимаемая информативность отзыва [298]

Перечисленные выше факторы представляют особый интерес для изучения в области управления и маркетинга, поскольку помогут принимать более эффективные решения по управлению продуктом и брендом, а также позволят лучше понимать потребителя. В настоящий момент было проведено разведывательное эмпирическое исследование по изучению покупателей в электронной коммерции и их отношения к написанию / чтению отзывов. На следующем шаге предпринимается попытка систематизации знания об использовании методов сетевого анализа для изучения отзывов, оставленных в сети Интернет.

### **13.5.2 Разведывательное эмпирическое исследование**

Данное исследование ставило перед собой задачу более детального изучения покупателей молодого возраста (18-25 лет) г. Москвы, как одной из лидирующих социально-демографических групп, совершающих онлайн-покупки, и их поведения относительно написания отзывов о покупке в онлайне. В частности, в данном исследовании была предпринята попытка дать ответы на следующие исследовательские вопросы:

- Есть ли связь между частотой и длиной оставляемого отзыва?
- Как можно классифицировать потребителей в возрасте 18-25 лет, исходя из их особенностей онлайн-поведения?
- Что мотивирует молодых людей оставлять отзывы о продукте в интернете?

Исследование строится на смешанной методологии с использованием количественных данных, собранных с 2017 по 2022гг. среди слушателей онлайн-курса «Маркетинг», которое размещено

на платформе «Открытое образование», а также качественных данных, полученных в 2022 году. Эмпирической базой количественного этапа исследования стали 1254 респондента (в возрасте от 18 до 25 лет, проживающих в городе Москва), а также 16 информантов, отобранных по невероятностной выборке с заданными критериями. Важно отметить, что в анализируемой эмпирической базе большая часть выборки приходится на 2018-2021 гг. (88%).

В целом, в проанализированной выборке 48% респондентов совсем не оставляют отзывы о компаниях/брендах/продуктах в интернете (на форумах, в социальных сетях, на сайтах компаний, специализированных порталах и др.), 17% делают это один раз в год или реже. Однако 56% опрошенных молодых людей говорят, что они принимают решения о покупке на основе обзоров и отзывов. Одновременно с этим растёт доля тех, кто стал чаще оставлять отзывы в интернете (на основе самоопределения респондентов): так, в 2017 году 56% скорее не согласились, не согласились, полностью не согласились с высказыванием «За последний год я стал чаще оставлять отзывы в интернете», в то время как в 2021-2022 годах доля таких респондентов стала меньше 50% (см. рис. 1).

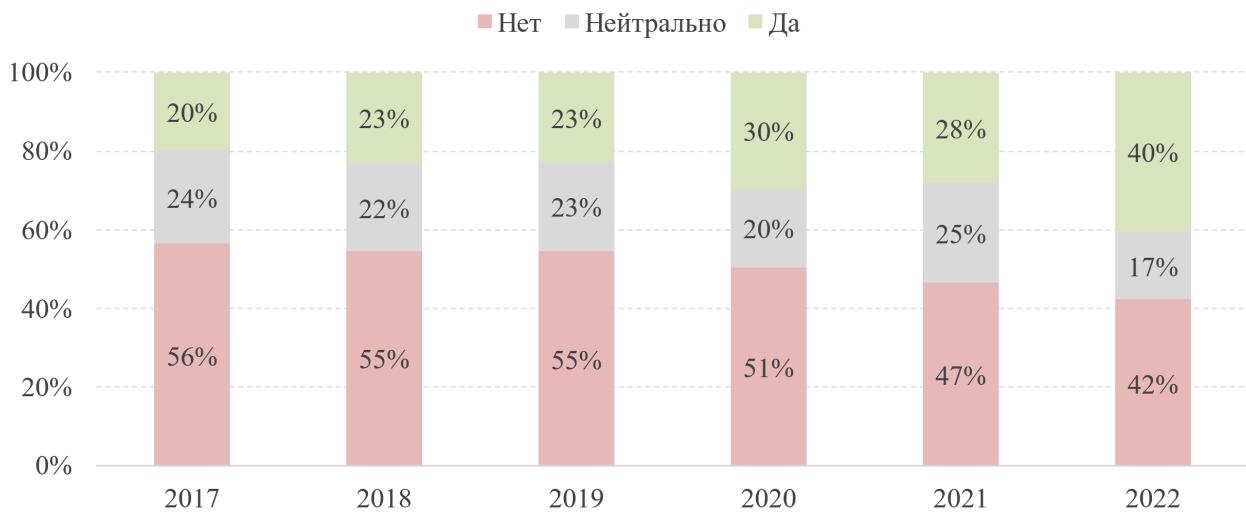


Рис. 1. Укрупнённое распределение ответов на вопрос-высказывание «За последний год я стал чаще оставлять отзывы в интернете»

Далее опишем основные поведенческие характеристики респондентов в возрасте 18-25 лет и проживающих в г.Москве, кто оставляет отзывы в интернете ( $n = 656$  респондентов). Среди тех, кто оставляет отзывы, 92% готовы написать как положительный, так и отрицательный отзыв. Это свидетельствует о том, что стереотип о написании только негативных отзывов среди молодой аудитории Москвы не распространён. Также большинство из них (66%) склонны доверять информации потребителей о компаниях, продуктах, товарах, оставленной в интернете.

В результате проведённого корреляционного анализа была обнаружена умеренная связь между частотой и длиной написанных отзывов. Так, если пользователи чаще оставляют отзывы, то они их пишут в более развёрнутом формате. Критерий Хи-квадрат, применённый к анализу таблиц сопряжённости, показал, что оценочность содержания отзыва (положительный или отрицательный характер отзыва) не связан с его длиной. Таким образом, негативный отзыв о товаре и, возможно, негативный опыт совершения покупки необязательно связан с тем, что пользователь оставит длинный отзыв. Таким образом, важно понимать, что мотивирует и демотивирует пользователей оставлять отзывы.

Перед тем, как классифицировать молодых людей на несколько групп на основе их покупательского

поведения, доверию интернет-торговле и поведения, связанного с написанием отзывов, на первом этапе проводился факторный анализ, чтобы уменьшить количество анализируемых высказываний и отобрать латентные переменные. В результате было выявлено 5 факторов: (1) отношение к онлайн покупкам в целом (2) лояльность к интернет-магазинам (3) безопасность покупок в интернет-магазинах (4) частота и длина отзывов (5) внимание к обзорам в онлайн-магазинах (см. табл. 1).

Табл. 1. Факторные нагрузки (5 компонент)

Высказывание	1	2	3	4	5
Следующую покупку я совершу в интернет-магазине, где уже покупал раньше					.816
Я совершу покупку в интернет-магазине, в котором уже покупал раньше, в течение следующего года					.809
Я намерен продолжать совершать покупки в интернет-магазине, в котором уже покупал раньше, и не собираюсь от этого отказываться					.795
Если бы мне пришлось снова купить тот же товар, я бы купил его в том же интернет-магазине					.784
Я рекомендую своим друзьям / коллегам / родственникам интернет-магазины, в которых совершаю покупки					.750
Я рассказываю своим друзьям про интернет-магазины, где совершаю покупки					.744
Я бы хотел, чтобы мои друзья / коллеги / родственники покупали в тех же интернет-магазинах, где покупаю я					.628
Мне легко взаимодействовать с интернет-магазином во время совершения покупки					.781
В целом, совершать покупки в интернете легко					.779
Мне легко совершать любые операции во время покупки в интернет-магазине					.764
В целом, в интернете удобно покупать					.742
Мне нравится совершать покупки в интернете					.733
Процесс покупки в интернете имеет много преимуществ					.724
Мне нравится искать товары в интернете					.657
Я беспокоюсь, что мои финансовые данные могут быть переданы другим компаниям без моего согласия					.904
Я беспокоюсь о сохранности персональных данных в интернете					.864
Я беспокоюсь о безопасности финансовых операций в интернете					.854
Я чувствую себя некомфортно, оставляя в интернете номер банковской карты					.849
Сообщения, отправленные через интернет, могут быть прочтены посторонними людьми или компаниями без моего ведома					.677
Я пишу длинные отзывы (более 1-2 предложений)					.896
За последний год я стал чаще оставлять отзывы в интернете					.889

Высказывание	1	2	3	4	5
Чем больше отзывов, тем выше вероятность, что я выберу товар и/или магазин для покупки					.866
Я обращаю внимание на количество отзывов при выборе товара и/или магазина					.862

Затем на основе результатов факторного анализа был проведён кластерный анализ. Было выявлено, что не все молодые люди г. Москвы в возрасте 18-25 лет имеют схожие паттерны поведения и отношения к онлайн-покупкам. В результате проведённого кластерного анализа были выявлены 5 кластеров. Так, например, кластер «Вдумчивые» склонны оставлять более длинные отзывы. Такие молодые люди в целом позитивно относятся к совершению покупок онлайн и доверяют отзывам, которые написаны в интернете. Одновременно с этим, среди кластера «Скептики» больше тех, кто не доверяет онлайн-транзакциям, в связи с чем реже совершает покупки в онлайне, реже оставляет отзывы и в целом более негативно относится к совершению онлайн-покупок.

Чтобы ответить на вопрос о том, как мотивировать молодёжь оставлять отзывы, были проведены глубинные интервью. В результате были выявлены такие драйверы, как:

1. недовольство товаром (в явной критической степени), в связи с чем покупатели хотят поделиться негативным опытом с другими потенциальными покупателями и предостеречь их;
2. полная удовлетворённость товаром (превосходящая ожидания);
3. получение материальных бонусов и стимулирующих действий от продавца. Среди основных барьеров были выявлены: отсутствие анонимности отзыва, отсутствие внутренней мотивации к оставлению отзыва, траты времени, а также отсутствие вознаграждение.

В результате проведённого эмпирического исследования, было изучено онлайн-поведение молодёжи г.Москвы в возрасте 18-25 лет. Были определены основные поведенческие паттерны по отношению к онлайн-покупкам и оставлению отзывов в онлайне. В рамках продолжения исследования планируется использовать такие методы как SNA и SEM для моделирования и определения влияния роли отзывов о совершении онлайн-покупок в интернете, а также более детального изучения текстов самих отзывов.

### **13.5.3 2. Применение методологии сетевого анализа для изучения пользовательского контента**

Сильный рост развития платформ социальных сетей привел к всплеску пользовательского контента (UGC). Как исследователи, так и практики обращаются к анализу социальных сетей (SNA) как к надежному методу расшифровки сложных закономерностей и взаимосвязей в этих больших наборах данных. Анализ литературы показал, что есть пул исследований, которые демонстрируют применимость и эффективность SNA при анализе пользовательского контента.

Анализ совокупности текстов пользователей в Интернете с помощью (SNA) начинается с преобразованием текстовых данных в сетевое представление, где узлы (nodes) представляют пользователей или фрагменты текста, а связи (links) представляют отношения или взаимодействия

[428]. Затем проводится анализ текстов или отношений, которые позволяют решить ряд управленческих задач в области бизнеса (в частности, в маркетинге). Именно о применении методологии сетевого анализа для таких задач пойдёт речь в данной главе.

#### **13.5.4 2.1 Расширенное знание о пользовательском контенте вокруг продукта и бренда**

Во-первых, анализ пользовательского контента с помощью SNA позволяет определять наиболее ключевых влиятельных лиц. Это может быть сеть вокруг бренда или продукта, которая включает как покупателей, так и в том числе бренды-конкуренты. Обнаруживание наиболее значимых лиц в сети позволит наладить сотрудничество, а взаимодействие с этими акторами может расширить охваты и повлиять на результаты маркетинговых кампаний.

Помимо обнаружения отдельных акторов, SNA помогает идентифицировать сообщества, что в свою очередь улучшает настройку таргетинга. Так, после выявления сообществ и групп, имеющих общие интересы или поведение, маркетологи могут адаптировать маркетинговые стратегии для конкретных сообществ, гарантируя, что контент и сообщения будут откликаться у отдельных членов этих групп.

#### **13.5.5 2.2 Позиция на рынке и отношения с конкурентами**

Когда компания только запускает новый продукт или услугу, особенно важно отслеживать информацию, которой делятся пользователи, поскольку это может повлиять на дальнейшую судьбу всего бренда [203]. Чёткое понимание обсуждаемых тем, вычленение достоинств и недостатков продукта, позволяет маркетологам оптимизировать маркетинговые кампании через понимание влияния пользователей друг на друга.

Немаловажно оценивать позицию бренда и продукта относительно конкурентов. Если в одну сеть добавить в том числе названия брендов, то можно проанализировать связи и взаимодействие между пользователями, которые обсуждают бренд в сравнении с конкурентами [162]. С помощью проведения такого анализа можно получить карту позиционирования бренда относительно конкурентов, увидеть конкурентные преимущества и скорректировать маркетинговые кампании.

#### **13.5.6 Маркетинг и реклама**

В маркетинге есть целый ряд метрик, которые чаще всего измеряются классическими онлайн-опросами или проведением интервью. Однако изучение текстов отзывов может воссоздать более объективно многие метрики у маркетологов может появиться более чёткое понимание и видение своего продукта. Например, можно измерять вовлечённость клиентов через анализ моделей взаимодействия пользователей с брендом или пользователей друг с другом через упоминание бренда [187]. Эта позволяет оценить эффективность маркетинговых усилий для дальнейшего улучшения и усовершенствования стратегий для повышения вовлечённости клиентов.

Другой немаловажный аспект – это сарафанное радио (word-of-mouth marketing). Пользователи любят рекомендовать понравившийся товар (или проводить антирекламу непонравившемуся товару). Изучая потоки и обмен информации между пользователями, можно идентифицировать пользователей, наиболее часто разделяющие положительные оценки о бренде и поощрить дальнейшее развитие сарафанного радио с помощью маркетинговой поддержки [151].

Изучение потребительского контента позволяет определить пользователей, которые с наибольшей вероятностью будут защищать бренд и в целом лояльны к нему [416]. Разрабатывая и внедряя программы по промотированию бренда через таких авторов постов, позволить продвигать бренд в социальных сетях более эффективно и с меньшим вложением затрат.

### 13.5.7 2.4 Клиентский маркетинг

На основе отзывов в онлайне можно воссоздать как действующую базу клиентов, так и потенциальную. С помощью SNA можно проводить сегментацию клиентов наиболее естественным образом [359]. Это позволяет адаптировать маркетинговые активности (через рекламные сообщения, акции и т.д.) к предпочтениям конкретных сегментов.

Помимо прочего, SNA может стать рабочей методологией в разработке стратегий удержания клиентов или улучшения сервиса и обслуживания. Так, с помощью текстового анализа отзывов можно понять негативные настроения и отрицательно окрашенные темы, что в дальнейшем можем позволить улучшить обслуживание, а также решить проблемы, быстро реагируя на отзывы взаимодействуя с пользователями. Отдельно можно выделить и обнаружение «центральных» акторов сети, через которых можно не только удерживать клиентов, но и влиять на базу и в целом охватывать потенциальных клиентов.

## 13.6 3. Процесс применения SNA для анализа пользовательских отзывов

SNA может быть ценным инструментом для анализа отзывов потребителей, рассматривая отношения между пользователями, оставляющими отзывы (в дальнейшем будем называть их рецензентами), или связи между продуктами и рецензентами как единую сеть [19]. Ниже рассмотрим пошаговую методологию по применению SNA для анализа пользовательских отзывов.

### 13.6.1 3.1 Подготовка массива данных

На первом этапе необходимо собрать данные (data scrapping) пользовательских отзывов с интернет-платформ (это могут быть публикации в социальных сетях, отзывы на сайтах маркетплейсов, обсуждение товаров на форумах и т.д.). Для использования SNA будет необходимо собрать корпус текстовых данных на основе отзывов.

Далее необходима предварительная обработка этих данных. Будет произведена очистка данных (удаление ненужной информации, работа с пропущенными данными и т.д.), а также произведена стандартизация текстов в единый формат. Для того, чтобы язык поддавался изучению, будет необходимо токенизировать текст с помощью стемминга и лемматизации.

Последним шагом будет необходимо построить сеть на основе имеющихся данных, где узлы представляют собой пользователей (рецензентов) или фрагменты текста, а связями являются отношения или взаимодействия между ними. Связи могут также быть построены на ответах, упоминаниях или взаимных упоминаниях брендов / продуктов или рецензентов.

## 13.7 3.2 Анализ основных метрик SNA

Использование атрибутов позволяет более комплексно подойти к анализу текстов. Так, добавлений атрибутов узлам (node attributes) и связям (edge attributes) помогут провести сравнительный анализ текстов отзывов по частоте оставления публикаций определённым автором или более детально изучить динамические тренды при добавлении дат создания. Также можно добавить характеристики текста для сети (например, количество слов в отзыве), чтобы проверить гипотезу об эффективности оставления отзыва в зависимости от длины сообщения.

Метрики центральности (centrality measures) позволяют идентифицировать наиболее влиятельных пользователей [328] обнаружение инфлюенсеров, влияющих на продажи того или иного продукта), а также выявление наиболее значимых фрагментов текстов отзыва. Помимо обнаружения индивидуальных центральностей важно изучение сообществ (community detection). Так, с помощью применения алгоритмов возможно выявить наиболее и наименее влиятельные группы сообществ [296], а также выявить тематические группы текстов.

### 13.7.1 3.3 Методы анализа данных: более детальный взгляд

В качестве первого метода планируется использовать content analysis, который позволит с помощью методов изучения и обработки естественного языку изучить содержание текстов. В частности, можно использовать выделение основных тем (topic modelling) и определение ключевых слов для определения преобладающих тем в корпусе.

Далее можно использовать анализ настроений (sentiment analysis), чтобы понять тональность сообщения, оставленного пользователем в сети. В частности, используются техники машинного обучения для классификации текстов на положительные, нейтральные и отрицательные фрагменты текста и сообщения в целом.

Для изучения трендов во времени используется временной анализ (temporal analysis), который позволяет изучить в динамике сеть взаимодействий во времени, определить различия в обнаруженных темах, а также сравнить настроения пользователей в разный период времени.

### 13.7.2 3.4 Визуализация и выводы

Самой наглядной демонстрацией выводов при изучении сети является визуализация, которую можно нарисовать с помощью программного обеспечения R, Rjek или Gephi. Визуализация (в том числе интерактивная) позволит наглядно обозначить выводы, полученные в результате применения техник и методов, обозначенных в подпунктах 2 и 3. Наконец, интерпретация и анализ технических методологических показателей, рассмотренных в данной главе, позволит маркетологам и управленцам в разработке контент-стратегий, управлении сообществами, а также контролировать и влиять на настроения пользователей относительно продуктов и услуг.

1. Simmel G. Soziologie.: Untersuchungen über die Formen der Vergesellschaftung. / G. Simmel, Leipzig: Duncker & Humblot, 1908. 4 c.
2. Social Mechanisms: An Analytical Approach to Social Theory под ред. P. Hedström, R. Swedberg, 1-е изд., Cambridge University Press, 1998.
3. Kalman Filtering and Neural Networks под ред. S. Haykin, 1-е изд., Wiley, 2001.

4. Qualitative Netzwerkanalyse под ред. B. Hollstein, F. Straus, Wiesbaden: VS Verlag für Sozialwissenschaften, 2006.
5. Qualitative Netzwerkanalyse под ред. B. Hollstein, F. Straus, Wiesbaden: VS Verlag für Sozialwissenschaften, 2006.
6. Закон Российской Федерации «Об обеспечении доступа к информации о деятельности судов в Российской Федерации» // 2008.
7. Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications под ред. D. Lusher, J. Koskinen, G. Robins, 1-е изд., Cambridge University Press, 2012.
8. Mixed Methods Social Networks Research: Design and Applications под ред. S. Domínguez, B. Hollstein, 1-е изд., Cambridge University Press, 2014.
9. Understanding Large Temporal Networks and Spatial Networks под ред. V. Batagelj [и др.], 1-е изд., Wiley, 2014.
10. Постановление Президиума Верховного Суда РФ от 27.09.2017 «Об утверждении Положения о порядке размещения текстов судебных актов на официальных сайтах Верховного Суда Российской Федерации, судов общей юрисдикции и арбитражных судов в информационно-телекоммуникационной сети «Интернет» // 2017.
11. Digital 2021: the latest insights into the «state of digital» - We Are Social UK // We Are Social [Электронный ресурс]. URL: <https://wearesocial.com/uk/blog/2021/01/digital-2021-the-latest-insights-into-the-state-of-digital/> (дата обращения: 27.11.2023).
12. Они уехали. Новая волна российской эмиграции // Социодиггер. 2023.
13. International Conference on Questionnaire Design, Development, Evaluation and Testing (QDET2) [Электронный ресурс]. URL: <https://ww2.amstat.org/meetings/qdet2/> (дата обращения: 27.11.2023).
14. «End-to-end construction of NLP knowledge graph» - Академия Google [Электронный ресурс]. URL: [https://scholar.google.com/scholar?hl=ru&as\\_sdt=0%2C5&q=%E2%80%98End-to-end+construction+of+NLP+knowledge+graph%E2%80%99+&btnG=](https://scholar.google.com/scholar?hl=ru&as_sdt=0%2C5&q=%E2%80%98End-to-end+construction+of+NLP+knowledge+graph%E2%80%99+&btnG=) (дата обращения: 29.11.2023).
15. ГАС РФ «Правосудие» [Электронный ресурс]. URL: <https://sudrf.ru> (дата обращения: 12.11.2023).
16. Алгоритм Света [Электронный ресурс]. URL: <https://readymag.com/algorithmsveta/algorithmsveta/> (дата обращения: 11.11.2023).
17. Цель под ключ - PBWM.RU [Электронный ресурс]. URL: <https://pbwm.ru/articles/tsel-pod-klyuch> (дата обращения: 25.11.2023).
18. Agrawal P., Garg V. K., Narayanan R. Link Label Prediction in Signed Social Networks KAUST Research Repository, 2013.C. 2591–2597.
19. Aiello L. M. [и др.]. Friendship prediction and homophily in social media // ACM Transactions on the Web (TWEB). 2012. № 2 (6). C. 1–33.
20. Airoldi E. M. [и др.]. Mixed Membership Stochastic Blockmodels Curran Associates, Inc., 2008.
21. Akbaritabar A. [и др.]. Italian sociologists: a community of disconnected groups // Scientometrics. 2020. № 3 (124). C. 2361–2382.
22. Alba R. D. A graph-theoretic definition of a sociometric clique // Journal of Mathematical Sociology. 1973. № 1 (3). C. 113–126.
23. Al-Balla H., Al-Dossari H., Chikh A. Using an Exponential Random Graph Model to Recommend Academic Collaborators // Information. 2019. № 6 (10). C. 220.

24. Albatineh A. N., Niewiadomska-Bugaj M., Mihalko D. On Similarity Indices and Correction for Chance Agreement // *Journal of Classification*. 2006. № 2 (23). C. 301–313.
25. Amati V., Schönenberger F., Snijders T. A. B. Estimation of Stochastic actor-oriented models for the evolution of networks by generalized method of moments.
26. Anastasi A. Evolving Concepts of Test Validation // *Annual Review of Psychology*. 1986. № 1 (37). C. 1–16.
27. Aria M., Cuccurullo C. bibliometrix: An R-tool for comprehensive science mapping analysis // *Journal of informetrics*. 2017. № 4 (11). C. 959–975.
28. Aronson B. Peer influence as a potential magnifier of ADHD diagnosis // *Social Science & Medicine*. 2016. (168). C. 111–119.
29. Arora A. S., Sanni S. A. Ten Years of «Social Media Marketing» Research in the Journal of Promotion Management: Research Synthesis, Emerging Themes, and New Directions // *Journal of Promotion Management*. 2019. № 4 (25). C. 476–499.
30. Arrigo F., Higham D. J. Sparse matrix computations for dynamic network centrality // *Applied Network Science*. 2017. № 1 (2). C. 17.
31. Ashforth B. E. Petty Tyranny in Organizations: A Preliminary Examination of Antecedents and Consequences // *Canadian Journal of Administrative Sciences / Revue Canadienne des Sciences de l'Administration*. 1997. № 2 (14). C. 126–140.
32. Association A. E. R., Association A. P., Education N. C. on M. in Standards for educational and psychological testing // 1985.
33. Barabási A.-L. Network science // *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 2013. № 1987 (371). C. 20120375.
34. Barger V., Labrecque L. An Integrated Marketing Communications Perspective on Social Media Metrics.
35. Bar-Hen A., Barbillon P., Donnet S. Block models for generalized multipartite networks: Applications in ecology and ethnobiology // *Statistical Modelling*. 2022. № 4 (22). C. 273–296.
36. Bartolucci F., Pandolfi S. An exact algorithm for time-dependent variational inference for the dynamic stochastic block model // *Pattern Recognition Letters*. 2020. (138). C. 362–369.
37. Batagelj V. On fractional approach to analysis of linked networks // *Scientometrics*. 2020. № 2 (123). C. 621–633.
38. Batagelj V., Ferligoj A., Doreian P. Direct and indirect methods for structural equivalence // *Social Networks*. 1992. № 1 (14). C. 63–90.
39. Batagelj V., Ferligoj A., Doreian P. Generalized Blockmodeling 1999.
40. Batagelj V., Ferligoj A., Doreian P. Indirect Blockmodeling of 3-Way Networks Berlin: Springer, 2007.C. 151–159.
41. Batagelj V., Ferligoj A., Squazzoni F. The emergence of a field: a network analysis of research on peer review // *Scientometrics*. 2017. № 1 (113). C. 503–532.
42. Beaton D. E. [и др.]. Guidelines for the process of cross-cultural adaptation of self-report measures // *Spine*. 2000. № 24 (25). C. 3186–3191.
43. Beatty P. C., Willis G. B. Research synthesis: The practice of cognitive interviewing // *Public opinion quarterly*. 2007. № 2 (71). C. 287–311.
44. Belsky J. Transition to parenthood // *Medical Aspects of Human Sexualit*. 1986. № 9 (20). C. 56–59.

45. Berardo R., Scholz J. T. Self-Organizing Policy Networks: Risk, Partner Selection, and Cooperation in Estuaries: SELF-ORGANIZING POLICY NETWORKS // American Journal of Political Science. 2010. № 3 (54). C. 632–649.
46. Bertrand M., Schoar A. The Role of Family in Family Firms // Journal of Economic Perspectives. 2006. № 2 (20). C. 73–96.
47. Besag J. On the Statistical Analysis of Dirty Pictures // Journal of the Royal Statistical Society. Series B (Methodological). 1986. № 3 (48). C. 259–302.
48. Blei D. M. Latent Dirichlet Allocation // Journal of Machine Learning Research. 2003. № 3. C. 993–1022.
49. Block P. [и др.]. Change we can believe in: Comparing longitudinal network models on consistency, interpretability and predictive power // Social Networks. 2018. (52). C. 180–191.
50. Blondel V. D. [и др.]. Fast unfolding of communities in large networks 2008.
51. Borgatti S. P. Social Network Analysis, Two-Mode Concepts in под ред. R. A. Meyers, New York, NY: Springer New York, 2009.C. 8279–8291.
52. Borgatti S. P., Everett M. G. Network analysis of 2-mode data // Social Networks. 1997. № 3 (19). C. 243–269.
53. Børsting C., Thomsen S. Foundation ownership, reputation, and labour // Oxford Review of Economic Policy. 2017. № 2 (33). C. 317–338.
54. Box-Steffensmeier J. M. [и др.]. Time Series Analysis for the Social Sciences / J. M. Box-Steffensmeier, J. R. Freeman, M. P. Hitt, J. C. W. Pevehouse, 1-е изд., Cambridge University Press, 2014.
55. Box-Steffensmeier J. M., Jones B. S. Event history modeling: A guide for social scientists / J. M. Box-Steffensmeier, B. S. Jones, Cambridge University Press, 2004.
56. Boyatzis R. E. The competent manager: A model for effective performance / R. E. Boyatzis, John Wiley & Sons, 1991.
57. Brusco M., Doreian P. A real-coded genetic algorithm for two-mode KL-means partitioning with application to homogeneity blockmodeling // Social Networks. 2015. (41). C. 26–35.
58. Burkart M., Miglietta S., Ostergaard C. Why Do Boards Exist? Governance Design in the Absence of Corporate Law // 2021.
59. Butts C. T. A Relational Event Framework for Social Action // Sociological Methodology. 2008. № 1 (38). C. 155–200.
60. Butts C. T. A Relational Event Framework for Social Action // Sociological Methodology. 2008. № 1 (38). C. 155–200.
61. Carley, K. M. Network Text Analysis: The Network Position of Concepts Routledge., 2020.C. 79–100.
62. Ceolodo G., Snijders T. A., Wit E. C. Stochastic Actor Oriented Model with Random Effects // arXiv preprint arXiv:2304.07312. 2023.
63. Cerinšek M., Batagelj V. Network analysis of Zentralblatt MATH data // Scientometrics. 2015. № 1 (102). C. 977–1001.
64. Chabert-Liddell S.-C. [и др.]. A stochastic block model approach for the analysis of multilevel networks: An application to the sociology of organizations // Computational Statistics & Data Analysis. 2021. (158). C. 107179.
65. Chabert-Liddell S.-C. Statistical Learning of Collections of Networks with Applications in Ecology and Sociology.

66. Chakraborty M., Byshkin M., Crestani F. Patent citation network analysis: A perspective from descriptive statistics and ERGMs // PLOS ONE. 2020. № 12 (15). C. e0241797.
67. Chen C. C., Tseng Y.-D. Quality evaluation of product reviews using an information quality framework // Decision Support Systems. 2011. № 4 (50). C. 755–768.
68. Chen H. [и др.]. Network dynamics in university-industry collaboration: a collaboration-knowledge dual-layer network perspective // Scientometrics. 2022. № 11 (127). C. 6637–6660.
69. Chen H. [и др.]. Network dynamics in university-industry collaboration: a collaboration-knowledge dual-layer network perspective // Scientometrics. 2022. № 11 (127). C. 6637–6660.
70. Chen X. [и др.]. Missing Traffic Data Imputation and Pattern Discovery with a Bayesian Augmented Tensor Factorization Model // Transportation Research Part C: Emerging Technologies. 2019. (104). C. 66–77.
71. Cheng C. [и др.]. Fused Matrix Factorization with Geographical and Social Influence in Location-based Social Networks 2012.C. 17–23.
72. Chiuso A., Pillonetto G. A Bayesian approach to sparse dynamic network identification // Automatica. 2012. № 8 (48). C. 1553–1565.
73. Cho H., Yu Y. Link prediction for Interdisciplinary Collaboration via Co-authorship Network 2018. № 25 (8).
74. Choudhary S. [и др.]. A Survey of Knowledge Graph Embedding and Their Applications // 2021.
75. Chow J. H., Kokotovic P. V. Time scale modeling of dynamic networks with sparse and weak connections Lecture Notes in Control and Information Sciences / под ред. P. V. Kokotovic, A. Bensoussan, G. L. Blankenship, Berlin/Heidelberg: Springer-Verlag, 1987.C. 310–353.
76. Chu S.-C., Kamal S. The effect of perceived blogger credibility and argument quality on message elaboration and brand attitudes: An exploratory study // Journal of interactive Advertising. 2008. № 2 (8). C. 26–37.
77. Chuan P. M. [и др.]. Link prediction in co-authorship Networks Based on Hybrid Content Similarity Metric // Applied Intelligence. 2018. № 8 (48). C. 2470–2486.
78. Cobo M. J. [и др.]. An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field // Journal of Informetrics. 2011. № 1 (5). C. 146–166.
79. Cocchia A. Smart and Digital City: A Systematic Literature Review под ред. R. P. Dameri, C. Rosenthal-Sabroux, Cham: Springer International Publishing, 2014.C. 13–43.
80. Coleman J. S. Social capital in the creation of human capital // American journal of sociology. 1988. (94). C. S95–S120.
81. Commission I. T. The ITC guidelines for translating and adapting tests // 2017.
82. Cooley C. H. Social organization: A study of the larger mind / C. H. Cooley, New York, NY: Charles Scribner's Sons, 1909. xvii, 426 c.
83. Corneli M., Latouche P., Rossi F. Block modelling in dynamic networks with non-homogeneous Poisson processes and exact ICL // Social Network Analysis and Mining. 2016. № 1 (6). C. 55.
84. Corneli M., Latouche P., Rossi F. Multiple change points detection and clustering in dynamic networks // Statistics and Computing. 2018. № 5 (28). C. 989–1007.
85. Coskun M., Koyutürk M. Link Prediction in Large Networks by Comparing the Global View of Nodes in the Network 2015.C. 485–492.
86. Costas J., Taheri A. «The Return of the Primal Father» in Postmodernity? A Lacanian Analysis of Authentic Leadership // Organization Studies. 2012. № 9 (33). C. 1195–1216.

87. Cote R. Dark Side Leaders: Are their Intentions Benign or Toxic? // Journal of Leadership, Accountability and Ethics. 2018. № 2 (15).
88. Cowan P. Individual and Family Life Transitions: A Proposal for a New Definition Hillsdale // NJ: Lawrence Erlbaum Associates, Inc. 1991.
89. Cugmas M., Ferligoj A., Kronegger L. The stability of co-authorship structures // Scientometrics. 2016. № 1 (106). C. 163–186.
90. Cugmas M., Žiberna A. Approaches to blockmodeling dynamic networks: A Monte Carlo simulation study // Social Networks. 2023. (73). C. 7–19.
91. Dao V. L., Bothorel C., Lenca P. Community structure: A comparative evaluation of community detection methods // Network Science. 2020. № 1 (8). C. 1–41.
92. Dashtipour P., Vidaillet B. Christophe Dejours' psychodynamic theory of work and its implications for leadership 2016.
93. Daud N. N. [и др.]. Applications of Link Prediction in Social Networks: A Review 2020. (166). C. 102716.
94. Daugherty T., Eastin M. S., Bright L. Exploring consumer motivations for creating user-generated content // Journal of interactive advertising. 2008. № 2 (8). C. 16–25.
95. Davis A., Gardner B. B., Gardner M. R. Deep South: A social anthropological study of caste and class / A. Davis, B. B. Gardner, M. R. Gardner, Univ of South Carolina Press, 2009.
96. De La Haye K. [и др.]. How physical activity shapes, and is shaped by, adolescent friendships // Social Science & Medicine. 2011. № 5 (73). C. 719–728.
97. De La Haye K. [и др.]. Smoking Diffusion through Networks of Diverse, Urban American Adolescents over the High School Period // Journal of Health and Social Behavior. 2019. № 3 (60). C. 362–376.
98. De Nooy W., Mrvar A., Batagelj V. Exploratory Social Network Analysis with Pajek / W. De Nooy, A. Mrvar, V. Batagelj, 1-е изд., Cambridge University Press, 2005.
99. De Silva C. S., Warusavitharana E. J., Ratnayake R. An examination of the temporal effects of environmental cues on pedestrians' feelings of safety // Computers, Environment and Urban Systems. 2017. (64). C. 266–274.
100. De Sola Pool I., Kochen M. Contacts and influence // Social Networks. 1978. № 1 (1). C. 5–51.
101. Deave T., Johnson D., Ingram J. Transition to parenthood: the needs of parents in pregnancy and early parenthood // BMC Pregnancy and Childbirth. 2008. № 1 (8). C. 30.
102. Diaz-Bone R. Gibt es eine qualitative Netzwerkanalyse? // Historical Social Research / Historische Sozialforschung. 2008. № 4 (126) (33). C. 311–343.
103. Dijkstra J. K. [и др.]. Testing Three Explanations of the Emergence of Weapon Carrying in Peer Context: The Roles of Aggression, Victimization, and the Social Network // Journal of Adolescent Health. 2012. № 4 (50). C. 371–376.
104. Dong Q. [и др.]. Incorporating Explicit Knowledge in Pre-trained Language Models for Passage Re-ranking // 2022.
105. Doreian P., Ferligoj A., Kronegger L. On the dynamics of national scientific systems: a reply // Quality & Quantity. 2011. № 5 (45). C. 1025–1029.
106. Doreian P., Mrvar A. A partitioning approach to structural balance // Social Networks. 1996. № 2 (18). C. 149–168.
107. Driver M. The Lack of Power or the Power of Lack in Leadership as a Discursively Constructed Identity // Organization Studies. 2013. № 3 (34). C. 407–422.

108. Dzansi J. Foundations and Investment Performance: The role of non-financial motives // Unpublished Working Paper. Jönköping International Business School. 2011.
109. Einarsen S., Aasland M. S., Skogstad A. Destructive leadership behaviour: A definition and conceptual model // The Leadership Quarterly. 2007. № 3 (18). C. 207–216.
110. Elias N. What is Sociology? / N. Elias, Hutchinson, 1978. 196 c.
111. eLibrary.ru Сравнение уровня публикаций российских ученых в базах данных Web of Science, Scopus и RSCI // eLibrary.ru [Электронный ресурс]. URL: [https://elibrary.ru/wos\\_scopus\\_rsci.asp](https://elibrary.ru/wos_scopus_rsci.asp)? (дата обращения: 30.08.2023).
112. Emelin D. [и др.]. Injecting Domain Knowledge in Language Models for Task-Oriented Dialogue Systems // 2022.
113. Emirbayer M., Goodwin J. Network analysis, culture, and the problem of agency // American journal of sociology. 1994. № 6 (99). C. 1411–1454.
114. Entsieh A. A., Hallström I. K. First-time parents' prenatal needs for early parenthood preparation-A systematic review and meta-synthesis of qualitative literature // Midwifery. 2016. (39). C. 1–11.
115. Epstein J., Santo R. M., Guillemin F. A review of guidelines for cross-cultural adaptation of questionnaires could not bring out a consensus // Journal of clinical epidemiology. 2015. № 4 (68). C. 435–441.
116. Esquivel A. V., Rosvall M. Compression of flow can reveal overlapping-module organization in networks // Physical Review X. 2011. № 2 (1). C. 021025.
117. Falicov C. J. Family transitions: Continuity and change over the life cycle // Guilford Press. 1991.
118. Fanelli D., Gläzel W. Bibliometric Evidence for a Hierarchy of the Sciences // PLoS ONE. 2013. № 6 (8). C. e66938.
119. Fang L. [и др.]. A Knowledge-Enriched Ensemble Method for Word Embedding and Multi-Sense Embedding // IEEE Transactions on Knowledge and Data Engineering. 2023. № 6 (35). C. 5534–5549.
120. Farasat A. [и др.]. Probabilistic graphical models in modern social network analysis 2015. № 1 (5). C. 1–18.
121. Ferligoj A. [и др.]. Scientific collaboration dynamics in a national scientific system // Scientometrics. 2015. № 3 (104). C. 985–1012.
122. Ferligoj A. [и др.]. Scientific collaboration dynamics in a national scientific system // Scientometrics. 2015. № 3 (104). C. 985–1012.
123. Fine G. A., Kleinman S. Network and meaning: An interactionist approach to structure // Symbolic interaction. 1983. № 1 (6). C. 97–110.
124. Fortunato S., Barthélémy M. Resolution limit in community detection // Proceedings of the national academy of sciences. 2007. № 1 (104). C. 36–41.
125. Fortunato S., Hric D. Community detection in networks: A user guide // Physics Reports. 2016. (659). C. 1–44.
126. Freeman L. C. The Sociological Concept of «Group»: An Empirical Test of Two Models // American Journal of Sociology. 1992. № 1 (98). C. 152–166.
127. Freeman L. C. The Development of Social Network Analysis: A Study in the Sociology of Science / L. C. Freeman, Empirical Press, 2004. 205 c.
128. Freeman L. C. The development of social network analysis—with an emphasis on recent events London: Sage London, 2011.C. 26–39.

129. Friedkin N. E. Structural cohesion and equivalence explanations of social homogeneity // Sociological Methods & Research. 1984. № 3 (12). C. 235–261.
130. Fronzetti Colladon A. The Semantic Brand Score // Journal of Business Research. 2018. (88). C. 150–160.
131. Fronzetti Colladon A., Grippa F. Brand Intelligence Analytics под ред. A. Przegalinska, F. Grippa, P. A. Gloor, Cham: Springer International Publishing, 2020.C. 125–141.
132. Fronzetti Colladon A., Grippa F., Innarella R. Studying the association of online brand importance with museum visitors: An application of the semantic brand score // Tourism Management Perspectives. 2020. (33). C. 100588.
133. Fuhse J., Fuhse J. Social Networks of Meaning and Communication / J. Fuhse, J. Fuhse, Oxford, New York: Oxford University Press, 2022. 344 c.
134. Fuhse J., Mützel S. Tackling connections, structure, and meaning in networks: quantitative and qualitative methods in sociological network research // Quality & quantity. 2011. (45). C. 1067–1089.
135. Funke T., Becker T. Stochastic block models: A comparison of variants and inference methods // PLOS ONE. 23 апр. 2019 г. № 4 (14). C. e0215296.
136. Gabriel Y. Psychoanalytic approaches to leadership London ; Thousand Oaks, Calif: SAGE, 2011.C. 393–405.
137. Garfield E. Citation indexing - its theory and application in science, technology, and humanities / E. Garfield, New York: Wiley, 1979. 274 c.
138. Genest C., MacKay J. The Joy of Copulas: Bivariate Distributions with Uniform Marginals // The American Statistician. 1986. № 4 (40). C. 280.
139. Gerber E. R., Wellens T. R. Perspectives on Pretesting : «Cognition» in the Cognitive Interview? // Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique. 1997. № 1 (55). C. 18–39.
140. Getoor L. [и др.]. Learning Probabilistic Models of Link Structure 2002. (3). C. 679–707.
141. Ghoorjian K. Graph Algorithms for Large-Scale and Dynamic Natural Language Processing 2019.
142. Ghose A., Ipeirotis P. G. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics // IEEE transactions on knowledge and data engineering. 2010. № 10 (23). C. 1498–1512.
143. Giddens A. The Constitution of Society: Outline of the Theory of Structuration / A. Giddens, Berkeley & Los Angeles: University of California Press, 1984. 448 c.
144. Gignoux J. [и др.]. Emergence and complex systems: The contribution of dynamic graph theory // Ecological Complexity. 2017. (31). C. 34–49.
145. Glynn C. [и др.]. Bayesian Analysis of Dynamic Linear Topic Models 2015.
146. Gorin S. V. [и др.]. The Russian Science Citation Index (RSCI): the First Three Years (2016-2018) // European Science Editing. 2020. (46). C. e51051.
147. Gou F., Wu J. Triad Link Prediction Method Based on the Evolutionary Analysis with IoT in Opportunistic Social Networks 2021. (181). C. 143–155.
148. Govaert G., Nadif M. Block clustering with Bernoulli mixture models: Comparison of different approaches // Computational Statistics & Data Analysis. 2008. № 6 (52). C. 3233–3245.
149. Granovetter M. Economic action and social structure: The problem of embeddedness Routledge, 2018.C. 22–45.
150. Granovetter M. S. The strength of weak ties // American journal of sociology. 1973. № 6 (78). C.

1360–1380.

151. Hambrick M. E., Pegoraro A. Social Sochi: Using social network analysis to investigate electronic word-of-mouth transmitted through social media communities // International Journal of Sport Management and Marketing. 2014. № 3-4 (15). C. 120–140.
152. Han X. [и др.]. PTR: Prompt Tuning with Rules for Text Classification // 2021.
153. Handcock M. S. Assessing Degeneracy in Statistical Models of Social Networks.
154. Hanneke S., Fu W., Xing E. P. Discrete temporal models of social networks // Electronic Journal of Statistics. 2010. № none (4).
155. Hansmann H., Thomsen S. The governance of foundation-owned firms // Journal of Legal Analysis. 2021. № 1 (13). C. 172–230.
156. Hartigan J. A., Wong M. A. Algorithm AS 136: A K-Means Clustering Algorithm // Journal of the Royal Statistical Society. Series C (Applied Statistics). 1979. № 1 (28). C. 100–108.
157. Harzing A.-W. Two New Kids on the Block: How do Crossref and Dimensions Compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science? // Scientometrics. 2019. № 1 (120). C. 341–349.
158. Hasan M. A., Zaki M. J. A Survey of Link Prediction in Social Networks под ред. C. C. Aggarwal, Boston, MA: Springer US, 2011.C. 243–275.
159. Haupt M. R. [и др.]. Characterizing Twitter User Topics and Communication Network Dynamics of the «Liberate» Movement During COVID-19 Using Unsupervised Machine Learning and Social Network Analysis // Online Social Networks and Media. 2021. (21). C. 100114.
160. Häussling R. Allocation to Social Positions in Class: Interactions and Relationships in First Grade School Classes and Their Consequences // Current Sociology. 2010. № 1 (58). C. 119–138.
161. He L. [и др.]. KLMo: Knowledge Graph Enhanced Pretrained Language Model with Fine-Grained Relationships под ред. M.-F. Moens [и др.], Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021.C. 4536–4542.
162. He Z., Zheng L., He S. A novel approach for product competitive analysis based on online reviews // Electronic Commerce Research. 2023. № 4 (23). C. 2259–2290.
163. Hennig-Thurau T. [и др.]. Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the internet? // Journal of interactive marketing. 2004. № 1 (18). C. 38–52.
164. Higgs M. The Good, the Bad and the Ugly: Leadership and Narcissism // Journal of Change Management. 2009. № 2 (9). C. 165–178.
165. Hilt E. When did ownership separate from control? Corporate governance in the early nineteenth century // The Journal of Economic History. 2008. № 3 (68). C. 645–685.
166. Hoff P. D. Multiplicative Latent Factor Models for Description and Prediction of Social Networks // Computational and Mathematical Organization Theory. 2009. № 4 (15). C. 261–272.
167. Hoff P. D., Raftery A. E., Handcock M. S. Latent Space Approaches to Social Network Analysis // Journal of the American Statistical Association. 2002. № 460 (97). C. 1090–1098.
168. Hoffmann T. [и др.]. Community detection in networks without observing edges // Science Advances. 2020. № 4 (6). C. eaav1478.
169. Hogan J., Holland B. Using theory to evaluate personality and job-performance relations: A socioanalytic perspective. // Journal of Applied Psychology. 2003. № 1 (88). C. 100–112.

170. Hogan R., Hogan J. Hogan Personality Inventory Manual / R. Hogan, J. Hogan, 3-е изд., Tulsa: Hogan Assessment Systems, 2007.
171. Hogan R., Hogan J. Hogan Development Survey Manual / R. Hogan, J. Hogan, Second Edition-е изд., Tulsa, OK: Hogan Assessment Systems, 2009. 199 с.
172. Holland P. W., Laskey K. B., Leinhardt S. Stochastic blockmodels: First steps // Social networks. 1983. № 2 (5). C. 109–137.
173. Holland P. W., Leinhardt S. Local structure in social networks // Sociological methodology. 1976. (7). C. 1–45.
174. Hollands R. G. Critical interventions into the corporate smart city // Cambridge Journal of Regions, Economy and Society. 2015. № 1 (8). C. 61–77.
175. Hollstein B. Qualitative approach London: Sage London, 2011.C. 404–416.
176. Hollstein B. Qualitative Approaches 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom: SAGE Publications Ltd, 2014.C. 404–416.
177. Holme P. Modern temporal network theory: a colloquium // The European Physical Journal B. 2015. № 9 (88). C. 234.
178. Homans G. C. The human group / G. C. Homans, Routledge, 2017.
179. Hopkins S. B., Steurer D. Efficient Bayesian Estimation from Few Samples: Community Detection and Related Problems Berkeley, CA: IEEE, 2017.C. 379–390.
180. Hou L., Liu Y., He X. Research on the Mechanism of Regional Innovation Network in Western China Based on ERGM: A Case Study of Chengdu-Chongqing Shuangcheng Economic Circle // Sustainability. 2023. № 10 (15). C. 7993.
181. Hu L. [и др.]. A Survey of Knowledge Enhanced Pre-trained Language Models // 2023.
182. Huang G. C. [и др.]. The Interplay of Friendship Networks and Social Networking Sites: Longitudinal Analysis of Selection and Influence Effects on Adolescent Smoking and Alcohol Use // American Journal of Public Health. 2014. № 8 (104). C. e51–e59.
183. Huang Z., Li X., Chen H. Link Prediction Approach to Collaborative Filtering Denver CO USA: ACM, 2005.C. 141–142.
184. Hummon N. P., Carley K. Social networks as normal science // Social Networks. 1993. № 1 (15). C. 71–106.
185. Ingold K., Fischer M. Drivers of collaboration to mitigate climate change: An illustration of Swiss climate policy over 15 years // Global Environmental Change. 2014. (24). C. 88–98.
186. Ingwersen P., Serrano-López A. E. Smart city research 1990–2016 // Scientometrics. 2018. № 2 (117). C. 1205–1236.
187. Isler Z., Kiygi-Calli M., El Oraiby M. Babbling through social media: A cross-country study mapping out social networks using eWOM intentions // Electronic Markets. 2023. № 1 (33). C. 54.
188. Javed M. A. [и др.]. Community detection in networks: A multidisciplinary review // Journal of Network and Computer Applications. 2018. (108). C. 87–111.
189. Ji S. [и др.]. A Survey on Knowledge Graphs: Representation, Acquisition and Applications // IEEE Transactions on Neural Networks and Learning Systems. 2022. № 2 (33). C. 494–514.
190. Jobe J. B., Tourangeau R., Smith A. F. Contributions of survey research to the understanding of memory // Applied Cognitive Psychology. 1993. № 7 (7). C. 567–584.

191. Jugert P., Leszczensky L., Pink S. The Effects of Ethnic Minority Adolescents' Ethnic Self-Identification on Friendship Selection // *Journal of Research on Adolescence*. 2018. № 2 (28). C. 379–395.
192. Kannan R., Vempala S., Vetta A. On clusterings: Good, bad and spectral // *Journal of the ACM (JACM)*. 2004. № 3 (51). C. 497–515.
193. Karrer B., Newman M. E. Stochastic blockmodels and community structure in networks // *Physical review E*. 2011. № 1 (83). C. 016107.
194. Kaya, Thomsen S. The Governance of Foundation-Owned Firms // 2022.
195. Kets de Vries M. F. R. *Prisoners of leadership* / M. F. R. Kets de Vries, New York: Wiley, 1989. 246 c.
196. Kets de Vries M. F. R. *Organizations on the couch: clinical perspectives on organizational behavior and change* / M. F. R. Kets de Vries, 1. ed-e изд., San Francisco, Calif.: Jossey-Bass, 1991. 408 c.
197. Kets de Vries M. F. R., Miller D. *The neurotic organization* / M. F. R. Kets de Vries, D. Miller, 1st ed-e изд., San Francisco: Jossey-Bass, 1984. 241 c.
198. Khandelwal U. [и др.]. Generalization through Memorization: Nearest Neighbor Language Models 2019.
199. Kim A. Qualitative network analysis in the strategy of mixing methods in the social sciences: a systematic literature review // *Sociology: methodology, methods, mathematical modeling (Sociology: 4M)*. 2022. № 53 (27). C. 83–116.
200. Kim A., Maltseva D. Stability evaluation of the Russian sociologists online community: 2011-2018 years 2021.
201. Kim A., Maltseva D. Qualitative social network analysis: studying the field through the bibliographic approach // *Quality & Quantity*. 2023. C. 1–27.
202. Kim B. [и др.]. A review of dynamic network models with latent variables // *Statistics surveys*. 2018. (12). C. 105–135.
203. Kim K., Lee W.-R., Altmann J. SNA-based innovation trend analysis in software service networks // *Electronic markets*. 2015. (25). C. 61–72.
204. Kim S.-M. [и др.]. Automatically assessing review helpfulness 2006.C. 423–430.
205. Kinne B. J. Network Dynamics and the Evolution of International Cooperation // *American Political Science Review*. 2013. № 4 (107). C. 766–785.
206. Kinne B. J. Dependent Diplomacy: Signaling, Strategy, and Prestige in the Diplomatic Network // *International Studies Quarterly*. 2014. № 2 (58). C. 247–259.
207. Kirimtak A. [и др.]. Future Trends and Current State of Smart City Concepts: A Survey // *IEEE Access*. 2020. (8). C. 86448–86467.
208. Kiuru N. [и др.]. Pressure to drink but not to smoke: Disentangling selection and socialization in adolescent peer networks and peer groups // *Journal of Adolescence*. 2010. № 6 (33). C. 801–812.
209. Kleinberg J. M. Hubs, authorities, and communities // *ACM Computing Surveys*. 1999. № 4es (31). C. 5.
210. Kochkarov A. A., Kochkarov R. A., Malinetskii G. G. Issues of dynamic graph theory // *Computational Mathematics and Mathematical Physics*. 2015. № 9 (55). C. 1590–1596.
211. Kohli C., Suri R., Kapoor A. Will social media kill branding? // *Business Horizons*. 2015. № 1 (58). C. 35–44.
212. Koller D., Friedman N. *Probabilistic Graphical Models: Principles and Techniques* / D. Koller, N. Friedman, Nachdr.-e изд., Cambridge, Mass.: MIT Press, 2010. 1231 c.
213. Kpiebaareh M. Y. [и др.]. A Generic Graph-Based Method for Flexible Aspect-Opinion Analysis of

- Complex Product Customer Feedback // Information. 2022. № 3 (13). C. 118.
214. Kretschmer D., Leszczensky L. In-Group Bias or Out-Group Reluctance? The Interplay of Gender and Religion in Creating Religious Friendship Segregation among Muslim Youth // Social Forces. 2022. № 3 (100). C. 1307–1332.
215. Krivitsky P. N. Exponential-family random graph models for valued networks // Electronic Journal of Statistics. 2012. № none (6).
216. Krivitsky P. N., Handcock M. S. A Separable Model for Dynamic Networks // Journal of the Royal Statistical Society Series B: Statistical Methodology. 2014. № 1 (76). C. 29–46.
217. Krivitsky P. N., Handcock M. S. A Separable Model for Dynamic Networks // Journal of the Royal Statistical Society Series B: Statistical Methodology. 2014. № 1 (76). C. 29–46.
218. Kronegger L. [и др.]. Collaboration structures in Slovenian scientific communities // Scientometrics. 2012. № 2 (90). C. 631–647.
219. Kronegger L. [и др.]. Collaboration structures in Slovenian scientific communities // Scientometrics. 2012. № 2 (90). C. 631–647.
220. Kroonenberg P. M. Applied Multiway Data Analysis / P. M. Kroonenberg, John Wiley & Sons, 2008. 615 c.
221. Kruschke J. K. Bayesian data analysis // WIREs Cognitive Science. 2010. № 5 (1). C. 658–676.
222. Kuo T.-T. [и др.]. Unsupervised Link Prediction using Aggregative Statistics on Heterogeneous Social Networks Chicago Illinois USA: ACM, 2013.C. 775–783.
223. Lapham S. C. [и др.]. How important is perception of safety to park use? A four-city survey // Urban Studies. 2016. № 12 (53). C. 2624–2636.
224. Latour B. Reassembling the Social: An Introduction to Actor-Network-Theory / B. Latour, OUP Oxford, 2007. 312 c.
225. Lauscher A. [и др.]. Specializing Unsupervised Pretraining Models for Word-Level Semantic Similarity под ред. D. Scott, N. Bel, C. Zong, Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020.C. 1371–1383.
226. Lee C., Wilkinson D. J. A review of stochastic block models and extensions for graph clustering // Applied Network Science. 2019. № 1 (4). C. 1–50.
227. Lee Y., Nelder J. A. Hierarchical Generalized Linear Models // Journal of the Royal Statistical Society: Series B (Methodological). 1996. № 4 (58). C. 619–656.
228. Leguia M. G. [и др.]. Reconstructing Dynamical Networks via Feature Ranking // Chaos: An Interdisciplinary Journal of Nonlinear Science. 2019. № 9 (29). C. 093107.
229. Leifeld P., Cranmer S. J. A theoretical and empirical comparison of the temporal exponential random graph model and the stochastic actor-oriented model // Network Science. 2019. № 1 (7). C. 20–51.
230. Leifeld P., Cranmer S. J. The stochastic actor-oriented model is a theory as much as it is a method and must be subject to theory tests // Network Science. 2022. № 1 (10). C. 15–19.
231. Leifeld P., Cranmer S. J., Desmarais B. A. Temporal Exponential Random Graph Models with **btergm** : Estimation and Bootstrap Confidence Intervals // Journal of Statistical Software. 2018. № 6 (83).
232. Leifeld P., Cranmer S. J., Desmarais B. A. Temporal Exponential Random Graph Models with **btergm** : Estimation and Bootstrap Confidence Intervals // Journal of Statistical Software. 2018. № 6 (83).
233. Lerner J. [и др.]. Conditional independence in dynamic networks // Journal of Mathematical Psychology.

2013. № 6 (57). C. 275–283.
234. Levine Y. [и др.]. SenseBERT: Driving Some Sense into BERT под ред. D. Jurafsky [и др.], Online: Association for Computational Linguistics, 2020.C. 4656–4667.
235. Li Q. [и др.]. OERL: Enhanced Representation Learning via Open Knowledge Graphs // IEEE Transactions on Knowledge and Data Engineering. 2023. № 9 (35). C. 8880–8892.
236. Li W.-J., Yeung D. Y., Zhang Z. Generalized Latent Factor Models for Social Network Analysis Barcelona, Spain;: 2011.C. 1705.
237. Liang X., Liu A. M. M. The evolution of government sponsored collaboration network and its impact on innovation: A bibliometric analysis in the Chinese solar PV sector // Research Policy. 2018. № 7 (47). C. 1295–1308.
238. Liben-Nowell D., Kleinberg J. The link-Prediction Problem for Social Networks // Journal of the American Society for Information Science and Technology. 2007. № 7 (58). C. 1019–1031.
239. Lipman-Blumen J. Toxic Leadership: A Conceptual Framework под ред. F. Bournois [и др.], London: Palgrave Macmillan UK, 2010.C. 214–220.
240. Liu W. [и др.]. K-BERT: Enabling Language Representation with Knowledge Graph // 2019.
241. Liu X. [и др.]. Network Embedding Based on a Quasi-Local Similarity Measure Lecture Notes in Computer Science / под ред. X. Geng, B.-H. Kang, Cham: Springer International Publishing, 2018.C. 429–440.
242. Liu Y. [и др.]. The Degree-Related Clustering Coefficient and its Application to Link Prediction // Physica A: Statistical Mechanics and Its Applications. 2016. (454). C. 24–33.
243. Lois D. Types of social networks and the transition to parenthood // Demographic Research. 2016. (34). C. 657–688.
244. Lorrain François, White Harrison C. Structural equivalence of individuals in social networks: The Journal of Mathematical Sociology: Vol 1, No 1 1971.
245. Lorrain F., White H. C. Structural equivalence of individuals in social networks // The Journal of mathematical sociology. 1971. № 1 (1). C. 49–80.
246. Lospinoso J., Snijders T. A. Goodness of fit for stochastic actor-oriented models // Methodological Innovations. 2019. № 3 (12). C. 205979911988428.
247. Lotker Z. Introduction to Evolving Social Networks Cham: Springer International Publishing, 2021.C. 167–185.
248. Lü L., Zhou T. Link Prediction in Complex Networks: A Survey // Physica A: Statistical Mechanics and its Applications. 2011. № 6 (390). C. 1150–1170.
249. Luce R. D., Perry A. D. A method of matrix analysis of group structure // Psychometrika. 1949. № 2 (14). C. 95–116.
250. Luhmann N. Social Systems / N. Luhmann, Stanford, Calif: Stanford University Press, 1995. 692 c.
251. Lundberg J. [и др.]. Collaboration Uncovered: Exploring the Adequacy of Measuring University-Industry Collaboration through Co-authorship and Funding // Scientometrics. 2006. № 3 (69). C. 575–589.
252. Lunguanu A., Contractor N. S. The Effects of Diversity and Network Ties on Innovations: The Emergence of a New Scientific Field // American Behavioral Scientist. 2015. № 5 (59). C. 548–564.
253. Ma K. [и др.]. Open Domain Question Answering with A Unified Knowledge Interface под ред. S. Muresan, P. Nakov, A. Villavicencio, Dublin, Ireland: Association for Computational Linguistics, 2022.C. 1605–1620.
254. Maccoby M. Narcissistic leaders: The incredible pros, the inevitable cons // Harvard Business Review.

2000. № 1 (78).

255. Mali F. [и др.]. Dynamic scientific co-authorship networks под ред. A. Scharnhorst, K. Börner, P. van den Besselaar, Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. C. 195–232.
256. Mallick C. [и др.]. Graph-Based Text Summarization Using Modified TextRank Advances in Intelligent Systems and Computing / под ред. J. Nayak [и др.], Singapore: Springer Singapore, 2019. C. 137–146.
257. Maltseva D., Batagelj V. Social network analysis as a field of invasions: bibliographic approach to study SNA development // Scientometrics. 2019. № 2 (121). C. 1085–1128.
258. Maltseva D., Batagelj V. Towards a systematic description of the field using keywords analysis: main topics in social networks // Scientometrics. 2020. № 1 (123). C. 357–382.
259. Maltseva D., Batagelj V. Towards a systematic description of the field using keywords analysis: main topics in social networks // Scientometrics. 2020. № 1 (123). C. 357–382.
260. Matias C., Miele V. Statistical Clustering of Temporal Networks Through a Dynamic Stochastic Block Model // Journal of the Royal Statistical Society Series B: Statistical Methodology. 2017. № 4 (79). C. 1119–1141.
261. Matias C., Rebafka T., Villers F. A semiparametric extension of the stochastic block model for longitudinal networks // Biometrika. 2018. № 3 (105). C. 665–680.
262. Matveeva N., Ferligoj A. Scientific collaboration in Russian universities before and after the excellence initiative Project 5-100 // Scientometrics. 2020. № 3 (124). C. 2383–2407.
263. Mead G. H. Mind, Self, and Society from the Standpoint of a Social Behaviorist / G. H. Mead, University of Chicago Press, 1967. 401 c.
264. Mehra A. [и др.]. The coevolution of friendship and leadership networks in small groups // Predator's game-changing designs: Research-based tools. 2009. C. 145–162.
265. Mercken L. [и др.]. Choosing adolescent smokers as friends: The role of parenting and parental smoking // Journal of Adolescence. 2013. № 2 (36). C. 383–392.
266. Meyer-Bäse A. [и др.]. Dynamical Graph Theory Networks Methods for the Analysis of Sparse Functional Connectivity Networks and for Determining Pinning Observability in Brain Networks // Frontiers in Computational Neuroscience. 2017. (11). C. 87.
267. Mirshahvalad A. [и др.]. Significant Communities in Large Sparse Networks // PLoS ONE. 2012. № 3 (7). C. e33721.
268. Mische A. Culture, Networks, and Interaction in Social Movement Publics // Simposio de Berlín. Marzo. 2008. (20).
269. Mitchell R. Web Scraping with Python: Collecting Data from the Modern Web / R. Mitchell, 1-е изд., O'Reilly Media, 2015. 253 c.
270. Moed H. F., Markusova V., Akoev M. Trends in Russian research output indexed in Scopus and Web of Science // Scientometrics. 2018. № 2 (116). C. 1153–1180.
271. Mohan A., Venkatesan R., Pramod K. V. A Scalable Method for Link Prediction in Large Real World Networks 2017. (109). C. 89–101.
272. Mohorko A., Hlebec V. Degree of cognitive interviewer involvement in questionnaire pretesting on trending survey modes // Computers in Human Behavior. 2016. (62). C. 79–89.
273. Mohrenberg S. Studying Policy Diffusion with Stochastic Actor-Oriented Models под ред. B. Hollstein, W. Matiaske, K.-U. Schnapp, Cham: Springer International Publishing, 2017. C. 163–188.

274. Mokken R. J. Cliques, clubs and clans // Quality & Quantity. 1979. № 2 (13). C. 161–173.
275. Molokwu B. C. Social Network Analysis: A Machine Learning Approach 2021.
276. Moradabadi B., Meybodi M. R. Link Prediction in Stochastic Social Networks: Learning Automata Approach 2018. (24). C. 313–328.
277. Morris M., Handcock M. S., Hunter D. R. Specification of Exponential-Family Random Graph Models: Terms and Computational Aspects // Journal of Statistical Software. 2008. № 4 (24).
278. Mørup M. Applications of tensor (multiway array) factorizations and decompositions in data mining // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2011. № 1 (1). C. 24–40.
279. Mørup M., Schmidt M. N. Bayesian Community Detection // Neural Computation. 2012. № 9 (24). C. 2434–2456.
280. Muniz C. P., Goldschmidt R., Choren R. Combining Contextual, Temporal and Topological Information for Unsupervised Link Prediction in Social Networks // Knowledge-Based Systems. 2018. (156). C. 129–137.
281. Murty S., Koh P. W., Liang P. ExpBERT: Representation Engineering with Natural Language Explanations под ред. D. Jurafsky [и др.], Online: Association for Computational Linguistics, 2020.C. 2106–2113.
282. Mydosh J. A. Spin glasses: an experimental introduction / J. A. Mydosh, CRC Press, 1993.
283. Nam T., Pardo T. A. Conceptualizing smart city with dimensions of technology, people, and institutions College Park Maryland USA: ACM, 2011.C. 282–291.
284. Nasiri E. [и др.]. Impact of Centrality Measures on the Common Neighbors in Link Prediction for Multiplex Networks // Big Data. 2022. (10). C. 138–150.
285. Newman M. E. Fast algorithm for detecting community structure in networks // Physical review E. 2004. № 6 (69). C. 066133.
286. Newman M. E. J. Spectral community detection in sparse networks 2013.
287. Newman M. E., Girvan M. Finding and evaluating community structure in networks // Physical review E. 2004. № 2 (69). C. 026113.
288. Nguyen G. H. [и др.]. Continuous-Time Dynamic Network Embeddings Lyon, France:, 2018.C. 969–976.
289. Nooy W. de, Mrvar A., Batagelj V. Exploratory Social Network Analysis with Pajek: Revised and Expanded Edition for Updated Software / W. de Nooy, A. Mrvar, V. Batagelj,.
290. Norris J. R. Markov chains / J. R. Norris, Cambridge university press, 1998.
291. Nowicki K., Snijders T. A. B. Estimation and Prediction for Stochastic Blockstructures // Journal of the American Statistical Association. 2001. № 455 (96). C. 1077–1087.
292. Obholzer A. Psychoanalytic contributions to authority and leadership issues // Leadership & Organization Development Journal. 1996. № 6 (17). C. 53–56.
293. Osman A. H., Barukub O. M. Graph-based text representation and matching: A review of the state of the art and future challenges // IEEE Access. 2020. (8). C. 87562–87583.
294. Özcan A., Öğüdücü Ş. G. Temporal Link Prediction Using Time Series of Quasi-Local Node Similarity Measures IEEE Computer Society, 2016.C. 381–386.
295. Pan Y., Fond M. Evaluating multilingual questionnaires: A sociolinguistic perspective 2014.C. 181–194.
296. Paranjape A., Benson A. R., Leskovec J. Motifs in temporal networks 2017.C. 601–610.
297. Paranyushkin D. InfraNodus: Generating Insight Using Text Network Analysis San Francisco CA USA: ACM, 2019.C. 3584–3589.
298. Park D.-H., Lee J. eWOM overload and its effect on consumer behavioral intention depending on consumer

- involvement // Electronic commerce research and applications. 2008. № 4 (7). C. 386–398.
299. Pas S. L. van der, Vaart A. van der Bayesian community detection 2018.
300. Pateman C. The Sexual Contract / C. Pateman, Stanford: Stanford University Press, 1988.
301. Pattison P., Robins G. 9. Neighborhood-Based Models for Social Networks // Sociological Methodology. 2002. № 1 (32). C. 301–337.
302. Peixoto T. P. Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models // Physical Review E. 2014. № 1 (89). C. 012804.
303. Peixoto T. P. Bayesian Stochastic Blockmodeling под ред. P. Doreian, V. Batagelj, A. Ferligoj, Wiley, 2019. C. 289–332.
304. Peixoto T. P., Rosvall M. Modelling sequences and temporal networks with dynamic community structures // Nature Communications. 2017. № 1 (8). C. 582.
305. Perianes-Rodriguez A., Waltman L., Van Eck N. J. Constructing Bibliometric Networks: A Comparison Between Full and Fractional Counting // Journal of Informetrics. 2016. № 4 (10). C. 1178–1195.
306. Persson C. [и др.]. Maps of sparse Markov chains efficiently reveal community structure in network flows with memory // arXiv preprint arXiv:1606.08328. 2016.
307. Peters K. [и др.]. Social Media Metrics — A Framework and Guidelines for Managing Social Media // Journal of Interactive Marketing. 2013. № 4 (27). C. 281–298.
308. Pike T. W. Collaboration Networks and Scientific Impact among Behavioral Ecologists 2010. № 2 (21). C. 431–435.
309. Pittel B., Spencer J., Wormald N. Sudden Emergence of a Giantk-Core in a Random Graph // Journal of Combinatorial Theory, Series B. 1996. № 1 (67). C. 111–151.
310. Pons P., Latapy M. Computing communities in large networks using random walks Springer, 2005. C. 284–293.
311. Portes A. Social capital: Its origins and applications in modern sociology // Annual review of sociology. 1998. № 1 (24). C. 1–24.
312. Portes A. The Two Meanings of Social Capital // Sociological Forum. 2000. № 1 (15). C. 1–12.
313. Pranckutė R. Web of Science (WoS) and Scopus: The Titans of Bibliographic Information in Today's Academic World 2021. C. 1–59.
314. Price D. J. D. S. Little Science, Big Science / D. J. D. S. Price, Columbia University Press, 1963.
315. Psorakis I., Roberts S., Sheldon B. Efficient Bayesian Community Detection using Non-negative Matrix Factorisation 2010.
316. Qin Y. [и др.]. ERICA: Improving Entity and Relation Understanding for Pre-trained Language Models via Contrastive Learning под ред. C. Zong [и др.], Online: Association for Computational Linguistics, 2021. C. 3350–3363.
317. Rakshit S. [и др.]. An integrated social network marketing metric for business-to-business SMEs // Journal of Business Research. 2022. (150). C. 73–88.
318. Rand W. M. Objective Criteria for the Evaluation of Clustering Methods // Journal of the American Statistical Association. 1971. № 336 (66). C. 846–850.
319. Redhead D., Von Rueden C. R. Coalitions and conflict: A longitudinal analysis of men's politics // Evolutionary Human Sciences. 2021. (3). C. e31.
320. Reichardt J., Bornholdt S. Statistical mechanics of community detection // Physical Review E. 2006. №

- 1 (74). C. 016110.
321. Rettinger A. [и др.]. Context-Aware Tensor Decomposition for Relation Prediction in Social Networks 2012. № 4 (2). C. 373–385.
322. Rhue L., Sundararajan A. Digital access, political networks and the diffusion of democracy // Social Networks. 2014. (36). C. 40–53.
323. Ricciardi F., Za S. Smart City Research as an Interdisciplinary Crossroads: A Challenge for Management and Organization Studies под ред. L. Mola, F. Pennarola, S. Za, Cham: Springer International Publishing, 2015.C. 163–171.
324. Richardson M., Domingos P. Markov Logic Networks 2006. № 1-2 (62). C. 107–136.
325. Ripley R. M. [и др.]. Manual for SIENA version 4.0 // University of Oxford. 2011.
326. Robins G., Pattison P., Woolcock J. Small and Other Worlds: Global Network Structures from Local Processes // American Journal of Sociology. 2005. № 4 (110). C. 894–936.
327. Roesler C., Broekel T. The role of universities in a network of subsidized R&D collaboration: The case of the biotechnology-industry in Germany // Review of Regional Research. 2017. № 2 (37). C. 135–160.
328. Romero D. M., Meeder B., Kleinberg J. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter 2011.C. 695–704.
329. Rosenthal S. A., Pittinsky T. L. Narcissistic leadership // The Leadership Quarterly. 2006. № 6 (17). C. 617–633.
330. Rossi A. [и др.]. Knowledge Graph Embedding for Link Prediction: A Comparative Analysis // ACM Transactions on Knowledge Discovery from Data. 2021. № 2 (15). C. 1–49.
331. Rosvall M. [и др.]. Different Approaches to Community Detection под ред. P. Doreian, V. Batagelj, A. Ferligoj, Wiley, 2019.C. 105–119.
332. Rosvall M., Bergstrom C. T. Maps of random walks on complex networks reveal community structure // Proceedings of the national academy of sciences. 2008. № 4 (105). C. 1118–1123.
333. Rosvall M., Bergstrom C. T. Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems // PloS one. 2011. № 4 (6). C. e18209.
334. Ruan J., Zhang W. An Efficient Spectral Algorithm for Network Community Discovery and Its Applications to Biological and Social Networks Omaha, NE, USA: IEEE, 2007.C. 643–648.
335. Rueden L. von [и др.]. Informed Machine Learning -- A Taxonomy and Survey of Integrating Knowledge into Learning Systems // IEEE Transactions on Knowledge and Data Engineering. 2021. C. 1–1.
336. Ruiz-Pérez R., Delgado López-Cózar E., Jiménez-Contreras E. Spanish Personal Name Variations in National and International Biomedical Databases: Implications for Information Retrieval and Bibliometric Studies // Journal of the Medical Library Association: JMLA. 2002. № 4 (90). C. 411–430.
337. Sachan D. [и др.]. Do Syntax Trees Help Pre-trained Transformers Extract Information? под ред. P. Merlo, J. Tiedemann, R. Tsarfaty, Online: Association for Computational Linguistics, 2021.C. 2647–2661.
338. Saeed M. [и др.]. RuleBERT: Teaching Soft Rules to Pre-Trained Language Models под ред. M.-F. Moens [и др.], Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021.C. 1460–1476.
339. Schaefer D. R., Haas S. A., Bishop N. J. A dynamic model of US adolescents' smoking and friendship networks // American journal of public health. 2012. № 6 (102). C. e12–e18.
340. Scheffler M., Brunzel J. Destructive leadership in organizational research: a bibliometric approach //

- Scientometrics. 2020. № 1 (125). C. 755–775.
341. Scheidlinger S. The psychology of leadership revisited: An overview // Group. 1980. № 1 (4). C. 5–17.
342. Schenk S. Neu-oder Restrukturierung des Geschlechterverhältnisses in Ostdeutschland // Berliner Journal für Soziologie. 1995. № 4 (5). C. 475–488.
343. Schmidt B. Costs and benefits of friendly boards during mergers and acquisitions // Journal of Financial Economics. 2015. № 2 (117). C. 424–447.
344. Schröder D., Thomsen S. Foundation ownership and financial performance—a global analysis. Groningen: October, 2021.
345. Schwartz H. S. Narcissistic process and corporate decay: the theory of the organization ideal / H. S. Schwartz, New York: New York University Press, 1990. 151 c.
346. Schwartz J. E. An Examination of Concor and Related Methods for Blocking Sociometric Data // Sociological Methodology. 1977. (8). C. 255.
347. Schweinberger M., Snijders T. A. B. Settings in Social Networks: A Measurement Model // Sociological Methodology. 2003. № 1 (33). C. 307–341.
348. Schweinberger M., Snijders T. A. B. Settings in Social Networks: A Measurement Model // Sociological Methodology. 2003. № 1 (33). C. 307–341.
349. Seidman S. B. Network structure and minimum degree // Social networks. 1983. № 3 (5). C. 269–287.
350. Seidman S. B. Internal cohesion of LS sets in graphs // Social Networks. 1983. № 2 (5). C. 97–107.
351. Seidman S. B., Foster B. L. A graph-theoretic generalization of the clique concept // Journal of Mathematical sociology. 1978. № 1 (6). C. 139–154.
352. Shrum W., Genuth J., Chompalov I. Structures of Scientific Collaboration / W. Shrum, J. Genuth, I. Chompalov, The MIT Press, 2007. 296 c.
353. Simmel G. Conflict and the web of group affiliations / G. Simmel, Simon and Schuster, 2010.
354. Singh A., Humphries M. D. Finding communities in sparse networks // Scientific Reports. 2015. № 1 (5). C. 8828.
355. Singh V. K. [и др.]. The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis // Scientometrics. 2021. № 6 (126). C. 5113–5142.
356. SKOLKOVO Портрет владельца капитала в России 2020. Что думают россияне о благосостоянии и о богатых людях // Бизнес-школа СКОЛКОВО - бизнес-образование, бизнес-обучение в Москве [Электронный ресурс]. URL: <https://www.skolkovo.ru/expert-opinions/portret-vladelca-kapitala-v-rossii-2020-chto-dumayut-rossiyane-o-blagosostoyanii-i-o-bogatyh-lyudyah/> (дата обращения: 25.11.2023).
357. Škulj D., Žiberna A. Stochastic blockmodeling of linked networks // Social Networks. 2022. (70). C. 240–252.
358. Skyler J. C., Bruce A. D., Jason W. M. Inferential Network Analysis / J. C. Skyler, A. D. Bruce, W. M. Jason, Cambridge University Press, 2021. 314 c.
359. Slaninová K. [и др.]. User segmentation based on finding communities with similar behavior on the web site IEEE, 2010.C. 75–78.
360. Smith M., Sarabi Y., Christopoulos D. Understanding collaboration patterns on funded research projects: A network analysis // Network Science. 2023. № 1 (11). C. 143–173.
361. Smith M., Sarabi Y., Christopoulos D. Understanding collaboration patterns on funded research projects: A network analysis // Network Science. 2023. № 1 (11). C. 143–173.

362. Snijders T. A. Stochastic actor-oriented models for network change // Journal of mathematical sociology. 1996. № 1-2 (21). C. 149–172.
363. Snijders T. A. Stochastic actor-oriented models for network dynamics // Annual review of statistics and its application. 2017. (4). C. 343–363.
364. Snijders T. A. B. The Statistical Evaluation of Social Network Dynamics // Sociological Methodology. 2001. № 1 (31). C. 361–395.
365. Snijders T. A. B. Markov Chain Monte Carlo Estimation of Exponential Random Graph Models.
366. Snijders T. A. B., Nowicki K. Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure // Journal of Classification. 1997. № 1 (14). C. 75–100.
367. Snijders T. A., Steglich C., Schweinberger M. Modeling the co-evolution of networks and behavior // Longitudinal models in the behavioral and related sciences. 2007. № 4 (31). C. 41–71.
368. Snijders T. A., Van de Bunt G. G., Steglich C. E. Introduction to stochastic actor-based models for network dynamics // Social networks. 2010. № 1 (32). C. 44–60.
369. Snijders T., Steglich C., Schweinberger M. Modeling the Coevolution of Networks and Behavior под ред. K. V. Montfort, J. Oud, A. Satorra, Routledge, 2017.C. 41–71.
370. Sommer E., Gamper M. Beyond structural determinism: Advantages and challenges of qualitative social network analysis for studying social capital of migrants // Global Networks. 2021. № 3 (21). C. 608–625.
371. Souder D. [и др.]. A behavioral understanding of investment horizon and firm performance // Organization Science. 2016. № 5 (27). C. 1202–1218.
372. Srilatha P., Manjula R. Paper Similarity Index based Link Prediction Algorithms in Social Networks: A Survey 2016. № 2. C. 87–94.
373. Stadtfeld C. The Micro-Macro Link in Social Networks // 2018.
374. Stadtfeld C., Hollway J., Block P. Rejoinder: DyNAMs and the Grounds for Actor-oriented Network Event Models // Sociological Methodology. 2017. № 1 (47). C. 56–67.
375. Steglich C., Snijders T. A., Pearson M. Dynamic networks and behavior: Separating selection from influence // Sociological methodology. 2010. № 1 (40). C. 329–393.
376. Stokman F. N., Sprenger C. GRADAP: Graph definition and analysis package 1989.
377. Strauss D. On a General Class of Models for Interaction // SIAM Review. 1986. № 4 (28). C. 513–527.
378. Su L. [и др.]. Scientometric cognitive and evaluation on smart city related construction and building journals data // Scientometrics. 2015. № 1 (105). C. 449–470.
379. Sun T. [и др.]. CoLAKE: Contextualized Language and Knowledge Embedding под ред. D. Scott, N. Bel, C. Zong, Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020.C. 3660–3670.
380. Sun Y. [и др.]. JointLK: Joint Reasoning with Language Models and Knowledge Graphs for Commonsense Question Answering // 2022.
381. Swaminathan V. [и др.]. Branding in a Hyperconnected World: Refocusing Theories and Rethinking Boundaries // Journal of Marketing. 2020. № 2 (84). C. 24–46.
382. Tan Q., Liu N., Hu X. Deep Representation Learning for Social Network Analysis // Frontiers in Big Data. 2019. (2). C. 1–10.
383. Tepper B. J. Abusive Supervision in Work Organizations: Review, Synthesis, and Research Agenda // Journal of Management. 2007. № 3 (33). C. 261–289.
384. The Centers for Population Health and Health Disparities Evaluation Working Group, Okamoto J. Scientific

- collaboration and team science: a social network analysis of the centers for population health and health disparities // Translational Behavioral Medicine. 2015. № 1 (5). C. 12–23.
385. Theodoridis S. Machine learning: a Bayesian and optimization perspective / S. Theodoridis, Academic press, 2015.
386. Thomsen S. Corporate ownership by industrial foundations // European Journal of Law and Economics. 1999. № 2 (7). C. 117–137.
387. Thomsen S. Industrial foundations: The Danish model Washington, D.C.: Brookings Institution Press, 2016. C. 3–33.
388. Thomsen S. The Nordic corporate governance model // Management and Organization Review. 2016. № 1 (12). C. 189–204.
389. Thomsen S. Foundation Ownership and Firm Performance: A Review of the International Evidence Oxford University Press, 2018. C. 66–85.
390. Thomsen S. [и др.]. Industrial foundations as long-term owners // Corporate Governance: An International Review. 2018. № 3 (26). C. 180–196.
391. Thomsen S., Degn S. M. The Charters of Industrial Foundations // 2014.
392. Thomsen S., Kavadis N. Enterprise foundations: law, taxation, governance, and performance // Annals of Corporate Governance. 2022. № 4 (6). C. 227–333.
393. Thurner P. W., Binder M. European Union transgovernmental networks: The emergence of a new political space beyond the nation-state? // European Journal of Political Research. 2009. № 1 (48). C. 80–106.
394. Tian H. [и др.]. SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis под ред. D. Jurafsky [и др.], Online: Association for Computational Linguistics, 2020. C. 4067–4076.
395. Tibshirani R. Regression Shrinkage and Selection Via the Lasso // Journal of the Royal Statistical Society: Series B (Methodological). 1996. № 1 (58). C. 267–288.
396. Townsend A. M. SMART CITIES: Big Data, Civic Hackers, and the Quest for a New Utopia 2013.
397. Traag V. A., Waltman L., Van Eck N. J. From Louvain to Leiden: guaranteeing well-connected communities // Scientific Reports. 2019. № 1 (9). C. 5233.
398. Traag V. A., Waltman L., Van Eck N. J. From Louvain to Leiden: guaranteeing well-connected communities // Scientific Reports. 2019. № 1 (9). C. 5233.
399. Van De Bunt G. G., Van Duijn M. A. J., Snijders T. A. B. Friendship Networks Through Time: An Actor-Oriented Dynamic Statistical Network Model // Computational & Mathematical Organization Theory. 1999. № 2 (5). C. 167–192.
400. Van Eck N., Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping // scientometrics. 2010. № 2 (84). C. 523–538.
401. Vaswani A. [и др.]. Attention Is All You Need // 2023.
402. Vermond D. [и др.]. The evolution and co-evolution of a primary care cancer research network: From academic social connection to research collaboration // PLOS ONE. 2022. № 7 (17). C. e0272255.
403. Vishnivetskaya A., Alexandrova E. «Smart city» concept. Implementation practice // IOP Conference Series: Materials Science and Engineering. 2019. (497). C. 012019.
404. Wang B., Bu Y., Xu Y. A quantitative exploration on reasons for citing articles from the perspective of cited authors // Scientometrics. 2018. № 2 (116). C. 675–687.
405. Wang J. [и др.]. Knowledge Prompting in Pre-trained Language Model for Natural Language Under-

- standing под ред. Y. Goldberg, Z. Kozareva, Y. Zhang, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022. C. 3164–3177.
406. Wang J., Ma Y., Yuan Y. Towards Fast Evaluation of Unsupervised Link Prediction by Random Sampling Unobserved Links // 2021.
407. Wang P. [и др.]. Link Prediction in Social Networks: the State-of-the-Art // 2014.
408. Wang R. [и др.]. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters // 2020.
409. Wang S. [и др.]. Signed Network Embedding in Social Media Society for Industrial and Applied Mathematics, 2017. C. 327–335.
410. Wang X. [и др.]. KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation // 2020.
411. Wang Y., Kockelman K. M. A Poisson-lognormal conditional-autoregressive model for multivariate spatial analysis of pedestrian crash counts across neighborhoods // Accident Analysis & Prevention. 2013. (60). C. 71–84.
412. Warner W. L. Social anthropology and the modern community // American Journal of Sociology. 1941. № 6 (46). C. 785–796.
413. Wasserman S., Faust K. Social Network Analysis: Methods and Applications / S. Wasserman, K. Faust, 1-е изд., Cambridge University Press, 1994.
414. Wasserman S., Faust K. Social Network Analysis: Methods and Applications / S. Wasserman, K. Faust, 1-е изд., Cambridge University Press, 1994.
415. Wasserman S., Faust K. Social Network Analysis: Methods and Applications / S. Wasserman, K. Faust, Cambridge University Press, 1994. 852 c.
416. Wassouf W. N. [и др.]. Predictive analytics using big data for increased customer loyalty: Syriatel Telecom Company case study // Journal of Big Data. 2020. (7). C. 1–24.
417. Watts S. A., Zhang W. Capitalizing on content: Information adoption in two online communities // Journal of the association for information systems. 2008. № 2 (9). C. 3.
418. Weber M. Economy and society: an outline of interpretive sociology / M. Weber, под ред. G. Roth, C. Wittich, Berkeley: University of California Press, 1978. 2 c.
419. Whetsell T. A. Democratic governance and global science: A longitudinal analysis of the international research collaboration network // PLOS ONE. 2023. № 6 (18). C. e0287058.
420. White H. C. Identity and Control: How Social Formations Emerge / H. C. White, Princeton: Princeton University Press, 2008. 451 c.
421. Wilde G. Gesellschaftsvertrag – Geschlechtervertrag // Ludwig, Gundula/Sauer, Birgit/Wöhl, Stefanie (Eds.): Staat und Geschlecht. Grundlagen und aktuelle Herausforderungen feministischer Staatstheorie. Baden-Baden: Nomos. 2009. C. 31–36.
422. Wu Z. [и др.]. Improving Local Clustering based Top-L Link Prediction Methods via Asymmetric Link Clustering Information 2018. (492). C. 1859–1874.
423. Xu K. Stochastic Block Transition Models for Dynamic Networks PMLR, 2015. C. 1079–1087.
424. Xu K. S., Hero A. O. Dynamic Stochastic Blockmodels for Time-Evolving Social Networks // IEEE Journal of Selected Topics in Signal Processing. 2014. № 4 (8). C. 552–562.
425. Xu R., Wunsch D. C. Recent advances in cluster analysis // International Journal of Intelligent Computing and Cybernetics. 2008. № 4 (1). C. 484–508.

426. Xu Y. [и др.]. Human Parity on CommonsenseQA: Augmenting Self-Attention with External Attention // 2022.
427. Yang T. [и др.]. Detecting communities and their evolutions in dynamic social networks—a Bayesian approach // Machine Learning. 2011. № 2 (82). C. 157–189.
428. Yang Z. [и др.]. Xlnet: Generalized autoregressive pretraining for language understanding // Advances in neural information processing systems. 2019. (32).
429. Yao Y. [и др.]. Kformer: Knowledge Injection in Transformer Feed-Forward Layers // 2022.
430. Yu W. [и др.]. A Survey of Knowledge-enhanced Text Generation // ACM Computing Surveys. 2022. № 11s (54). C. 227:1–227:38.
431. Yuloskov A. [и др.]. Smart Cities in Russia: Current Situation and Insights for Future Development // Future Internet. 2021. № 10 (13). C. 252.
432. Zachlod C. [и др.]. Analytics of social media data – State of characteristics and application // Journal of Business Research. 2022. (144). C. 1064–1076.
433. Zalk M. H. W. V. [и др.]. Peer Contagion and Adolescent Depression: The Role of Failure Anticipation // Journal of Clinical Child & Adolescent Psychology. 2010. № 6 (39). C. 837–848.
434. Zeng S. [и др.]. Double Graph Based Reasoning for Document-level Relation Extraction // 2020.
435. Zhang C. Understanding scientific collaboration from the perspective of collaborators and their network structures Philadelphia, USA: iSchools, 2016.
436. Zhang F. [и др.]. «Perception bias»: Deciphering a mismatch between urban crime and perception of safety // Landscape and Urban Planning. 2021. (207). C. 104003.
437. Zhang T. [и др.]. DKPLM: Decomposable Knowledge-enhanced Pre-trained Language Model for Natural Language Understanding // 2022.
438. Zhang W. [и др.]. Recommendation System in Social Networks with Topical Attention and Probabilistic Matrix Factorization 2019. № 10 (14). C. e0223967.
439. Zhang X. [и др.]. GreaseLM: Graph REASoning Enhanced Language Models for Question Answering // 2022.
440. Zhang Y., Leslie J. B., Hannum K. M. Trouble Ahead: Derailment Is Alive and Well // Thunderbird International Business Review. 2013. № 1 (55). C. 95–102.
441. Zhu J., Liu W. A Tale of Two Databases: the Use of Web of Science and Scopus in Academic Papers // Scientometrics. 2020. № 1 (123). C. 321–335.
442. Zhu J., Liu W. A Tale of Two Databases: the Use of Web of Science and Scopus in Academic Papers // Scientometrics. 2020. № 1 (123). C. 321–335.
443. Žiberna A. Generalized blockmodeling of valued networks // Social Networks. 2007. № 1 (29). C. 105–126.
444. Žiberna A. Evaluation of direct and indirect blockmodeling of regular equivalence in valued networks by simulations // Advances in Methodology and Statistics. 2009. № 2 (6).
445. Žiberna A. Blockmodeling of multilevel networks // Social Networks. 2014. (39). C. 46–61.
446. Žiberna A. Blockmodeling Linked Networks под ред. P. Doreian, V. Batagelj, A. Ferligoj, Wiley, 2019.C. 267–287.
447. Žiberna A. k-means-based algorithm for blockmodeling linked networks // Social Networks. 2020. (61). C. 153–169.
448. Žiberna A. Comparing different methods for one-mode homogeneity blockmodeling according to structural

- equivalence on binary networks // Advances in Methodology and Statistics. 2021. № 1 (18).
449. Антонова К. А., Дубровина А. Д., Якимова О. А. Проблемы на новом месте: что обсуждают российские эмигранты в чатах взаимопомощи? // Социодиггер. 2023. № 3-4 (4). С. 59–68.
450. Барковец А. А. Функционирование российской модели эндаумент-фондов М.: Минэкономразвития РФ, 2013. С. 55–67.
451. Булычева Е. Е., Мальцева Д. В. Выделение актуальных тематик в социологии: взгляд сквозь призму анализа сети цитирований // The monitoring of public opinion economic&social changes. 2020. № 6.
452. Винер Б. Е., Дивисенко К. С. Когнитивная структура современной российской социологии по данным журнальных ссылок 2012. № 4 (15). С. 144–166.
453. Горшков М. К., Шереги Ф. Э. Прикладная социология: методология // М.: Институт социологии РАН. 2011.
454. Девятко И. Ф. Методы социологического исследования / И. Ф. Девятко, Екатеринбург: Издательство Уральского университета, 1998.
455. Докука С. В. Использование стохастических акторно-ориентированных моделей для анализа коэволюции сетей и поведения // Мониторинг общественного мнения: экономические и социальные перемены. 2021. № 2. С. 273–285.
456. Докука С. В., Валеева Д. Р. Статистические модели для анализа динамики социальных сетей в исследованиях образования // Вопросы образования. 2015. № 1. С. 201–213.
457. Евстратов А. Г. Релокация россиян в Армению в свете спецоперации РФ на Украине // Архонт. 2022. (3).
458. Еременко Г. О. Сравнение уровня публикаций российских ученых в базах данных Web of Science, Scopus и RSCI [Электронный ресурс]. URL: [https://elibrary.ru/wos\\_scopus\\_rsci.asp](https://elibrary.ru/wos_scopus_rsci.asp).
459. Жучкова С. В., Ротмистров А. Н. Автоматическое извлечение текстовых и числовых веб-данных для целей социальных наук 2020. № 50-51. С. 141–183.
460. Ким А. В. Качественный сетевой анализ в стратегии смешивания методов в социальных науках: систематический обзор литературы // Социология: методология, методы, математическое моделирование (Социология:4М). 2021. № 53. С. 83–116.
461. Клинов И. А. [и др.]. МПУ Сколково. Модели управления благотворительными фондами – бенефициарными собственниками бизнес-компаний. Москва, 2023.
462. Костенко Н. [и др.]. Российская ризома: социальный портрет новой эмиграции // Re: Russia. 2023.
463. Кохут Х. Анализ самости: Системный подход к лечению нарциссических нарушений личности / Пер. с англ. А.М. Боковикова / Х. Кохут, Москва: Когито-Центр, 2003.
464. Мальцева Д. Сетевой подход как феномен социологической теории // Социологические исследования. 2018. № 4. С. 3–14.
465. Мальцева Д. В. Реляционная социология: новый этап в развитии анализа социальных сетей или самостоятельное направление? // Мониторинг общественного мнения: экономические и социальные перемены. 2014. № 4 (122). С. 3–14.
466. Мальцева Д. В., Fiala D. Russian Publications in Web of Science: A Bibliometric Study // (препринт) COLLNET Journal of Scientometrics and Information Management. 2023.
467. Мальцева Д. В., Ващенко В. А., Капустина Л. В. Методология обработки библиографических данных на русском языке для построения сетей коллaborации 2022.

468. Мисютина В. [и др.]. WTC Сколково. Исследование владельцев капиталов России 2015. Москва, 2015.
469. Моисеев С.П., Мальцева Д.В. Отбор источников для систематического обзора литературы: сравнение экспертного и алгоритмического подходов // Социология: методология, методы, математическое моделирование (Социология:4М). 2018. № 47. С. 7–43.
470. Павлова И. А. ПОСТРОЕНИЕ КАРТЫ СОПРИСУТСТВИЯ КЛЮЧЕВЫХ СЛОВ ПО ТЕМЕ «КАПИТАЛ ЗДОРОВЬЯ» В ПРОГРАММЕ VOSVIEWER // Векторы благополучия. 2023. № 2 (49). С. 38–54.
471. Подольская А. П., Харламова Е. Е. Целевой капитал как источник финансирования некоммерческой организации // Финансовая аналитика: проблемы и решения. 2016. № 284 (2). С. 31–42.
472. Поздняков М. Л. Организационные и структурные ограничения при доступе к судебным актам судов общей юрисдикции. (Серия «Аналитические записки по проблемам правоприменения»). // 2012.
473. Ребязина В. А. [и др.]. Развитие электронной коммерции в России: влияние пандемии COVID-19 / В. А. Ребязина, Е. Р. Шарко, С. М. Березка, А. Г. Старков, Москва: ИД НИУ ВШЭ, 2022. 72 с.
474. Рогозин Д. М. Когнитивный анализ опросного инструмента // Социологический журнал. 2000. № 3-4. С. 018–070.
475. Рогозин, Д. М. Когнитивный анализ опросного инструмента / Рогозин, Д. М.;, 2002.
476. Савченко П. В., Федорова М. Н., Шлихтер А. А. Эндаумент как институт социальных инвестиций // Вестник Института экономики Российской академии наук. 2015. № 2. С. 52–63.
477. Сафонова М. А., Винер Б. Е. Сетевой анализ социтирований этнологических публикаций в российских периодических изданиях: предварительные результаты 2013. № 36. С. 140–176.
478. Селизарова В. Куда уехали россияне из-за частичной мобилизации и что планируют делать дальше // Forbes. 2023.
479. Соколов М. М. [и др.]. Интеллектуальный ландшафт и социальная структура локального академического сообщества (Случай петербургской социологии) // 2012.
480. Соколова С. Ю. Фонды целевого капитала в системе обеспечения конкурентоспособности некоммерческих организаций // Экономический журнал. 2011. № 24. С. 73–78.
481. Таворкин Е. П. Основы методики социологического исследования / Таворкин Е. П., НИЦ ИНФРА-М, 2021. 239 с.
482. Трофимова И. Н. Международное сотрудничество российских исследователей: текущие позиции и тенденции: по данным Web of Science за 2018–2022 гг. 2023. № 4 (32). С. 178–198.
483. Хорни К. Невроз и личностный рост. Борьба за самореализацию / Пер. с англ. Е.И. Замфир / К. Хорни, Питер, 2019.
484. Ширманова И. 800 тысяч россиян могли покинуть страну в 2022 году // Если быть точным. 2023.