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

МЕТОДОЛОГИЧЕСКИЕ РАЗРАБОТКИ И ПРАКТИЧЕСКОЕ ПРИМЕНЕНИЕ

November 28, 2023

Содержание

1 Основные термины

1.1 Социальный капитал

Вот цитирование (**larranaga2013?**) и без скобок (**larranaga2013?**) — и сноска[1]. А вот пример ссылки на часть текста sec. 3.1.

Попробуем написать что-то про социальный капитал. Но никто из моих цитирований про это не писал. У

[1] Внутренняя сноска

меня есть ссылка на две статьи по блокмоделингу: такую (Batagelj et al., 1999) и такую Marcot & Penman (2019). Но эти статьи не подходят и для следующего раздела sec. 1.2. Еще что-то написали. ### Bridging ### Bonding

1.2 Социальное доверие

Также хотелось сделать картинку, но пока оставляю вот эту:

Модульный шрифт Джозефа Альберса

Figure 1: Модульный шрифт Джозефа Альберса

> Цитата - это важно. *А еще лучше писать ее курсивом.*

1.3 Актуальность

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?**).

1.4 Литература

1.5 Сети международной торговли

1.5.1 Введение

> Уолтер Рэли (1552–1618), английский мореплаватель, историк, поэт. *Кто владеет* ***мировой торговлей****, владеет богатствами мира, а следовательно – и самим миром.*

Рассказываем про сети международной торговли, как они образуются, что влияет на их развитие и т.п.

1.5.2 Графики

Как отмечается в (**bhattacharyaInternationalTradeNetwork2008a?**) торговые сети характеризуются высокой плотностью.[2] Вместе с тем, в работе (**deandradeUseNodesAttributes2018?**) отмечается сильная взаимозависимость.

Торговая сеть в 2010 году по наиболее крупному партнеру

Figure 2: Торговая сеть в 2010 году по наиболее крупному партнеру

Заключение

[2] 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 коллаборативности.

1.6 Experiments with model's robustity

Я читала такую статью: (Zhou et al., 2021), она написана Zhou et al. (2021). Надеюсь этот файлик загрузится ^^

1.7 1.5 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 экспертного 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 (Haupt et al., 2021). 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 мечеными узлами. Как утверждают Tan и соавторы (Tan et al., 2019): «Существующие методы можно разделить на две категории, например, методы, основанные на случайном 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 (Tan et al., 2019) 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 (Molokwu, 2021), «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 скудны 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 (Molokwu, 2021).

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 (P. Wang et al., 2014) 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 (Huang et al., 2005). Эффективное применение анализа сетевых связей также часто встречается в научных работах, посвященных сетям соавторства (Cho & Yu, 2018; Liben-Nowell & Kleinberg, 2007), в результате чего эта область является одной из наиболее процветающих в последние годы, поскольку способствует развитию эффективной системы совместной работы (Chuan et al., 2018; Huang et al., 2005). 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 (Daud et al., 2020). Социальные сети – это динамические образования (Leguia et al., 2019), 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. (Hasan & Zaki, 2011). В предыдущие годы появилось множество научных работ, посвященных острым вопросам предсказания связей, касающихся метрик, используемых для целей предсказания (Cho & Yu, 2018; Mohan et al., 2017), обсуждающих задачу предсказания в различных типах сетей (Gou & Wu, 2021; Nasiri et al., 2022). Наконец, появились обзоры (Daud et al., 2020; Hasan & Zaki, 2011; P. Wang et al., 2014), 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 (Daud et al., 2020), которая помогает структурировать обзор.

Одной из наиболее простых в применении и традиционных групп подходов являются подходы, основанные на сходстве (Daud et al., 2020; Moradabadi & Meybodi, 2018). 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 (Daud et al., 2020; Lü & Zhou, 2011), 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 (Daud et al., 2020; X. Liu et al., 2018), учитывают больше информации о соседних узлах (Lü & Zhou, 2011) 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 (X. Liu et al., 2018; Özcan & Öğüdücü, 2016; Srilatha & Manjula, 2016).

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 (Hasan & Zaki, 2011), так как упускаются потенциальные связи. Основные работы в области локально-индексных подходов ведутся в области совершенствования метрик вычисления сходства и алгоритмов вычисления сходства (Daud et al., 2020; Z. Wu et al., 2018). 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 (Coskun & Koyutürk, 2015; Muniz et al., 2018). 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 (Y. Liu et al., 2016; Özcan & Öğüdücü, 2016; S. Wang et al., 2017). 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 (Farasat et al., 2015). В этой группе методов исключительно важно обращать внимание на тип сети, так как в марковских и байесовских сетях используются разные методы и расчеты вероятностей (Farasat et al., 2015). Следует внимательно относиться к типу переменных, взаимности связей, типу взаимных связей и т.д. (Koller & Friedman, 2010). 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 (Daud et al., 2020), мы также можем структурировать наш обзор, опираясь на четыре типа подходов в группе вероятностных методов:

1. Модель тензорной факторизации вероятностей (Cheng et al., 2012; J. Wang et al., 2021). Как указано в (Daud et al., 2020), они являются логическими расширениями моделей Probability Matrix Factorization, используемых для решения задачи тензорной факторизации. Развитие этого подхода можно увидеть в работах (Chen et al., 2019; Rettinger et al., 2012; Zhang et al., 2019).

2. Модель вероятностных латентных переменных (Hoff, 2009; Li et al., 2011). 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 (Daud et al., 2020). 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 (Agrawal et al., 2013). Эта группа моделей применима только к сетям с подписями и решает задачу предсказания меток связей. Подписанные связи несут дополнительную информацию, уникальную для наблюдаемой 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 холодности. 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 (Daud et al., 2020).

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 (Daud et al., 2020).

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 (Daud et al., 2020), 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 коллаборативную фильтрацию (Daud et al., 2020), 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. Перцептрон 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) (Getoor et al., 2002) и марковские логические сети (MLN) (Richardson & Domingos, 2006). Неподконтрольный механизм, использующий статистику, был реализован Kyo и др (Kuo et al., 2013). 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 (Liben-Nowell & Kleinberg, 2007), также изучались Мунисом и др (Muniz et al., 2018) 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 (Liben-Nowell & Kleinberg, 2007): 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 (Nguyen et al., 2018). 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 (Nguyen et al., 2018) 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 (J. Wang et al., 2021) рассматривают проблему оценки предсказания связей без

наблюдения. Основная проблема оценки ненаблюдаемого предсказания связей заключается в том, что ненаблюдаемых 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 ненаблюдаемых 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, управляющих 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.

2 1.2 Акторно-ориентированные стохастические модели для изучения сетевой динамики и социального влияния

2.1 Введение

В контексте развития сетевого анализа исследования сетей в динамике становится 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 линзы, 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. ## 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. ### Применения

Дружба и влияние сверстников Применения SAM породило довольно много исследований на более маленьких выборках сетей дружбы, в которых исследовались реципрокность и гомофилия ка каналы социального влияния на курение, потребление веществ, ожирение и т. д. в группах сверстников. Среди подобных исследований: эффекты пола на распределение индивидуальных характеристик в сетях дружбы Van De Bunt и др., 1999]](https://www.zotero.org/google-docs/?gN9B2g); актуальные проблемы курения и употребления алкоголя через призму гомофилии: влияния сверстников [De La Haye и др., 2019; Huang и др., 2014; Kiuru и др., 2010; Schaefer и др., 2012] и родительского примера [Mercken и др., 2013].

Кроме того, на пересечении медицинской социологии и сетевого анализа, существуют также исследования, применяющие SAOM, для выявления социального влияния на вероятность заболеваний, например, СДВГ [Aronson, 2016], подростковой депрессии, [Zalk и др., 2010], образа жизни и ожирения [De La Haye и др., 2011]. В этой области также существуют работы о социальном влиянии этнического самоопределения [Jugert, Leszczensky, Pink, 2018], лидерских динамик [Mehra и др., 2009], религии [Kretschmer, Leszczensky, 2022] и владения оружием [Dijkstra и др., 2012] 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 [Ferligoj и др., 2015], а также провели дополнительную работу на лучших данных, указав на важность институциональных контекстов на среду

работы ученых, его не-механическую природу [Kronegger и др., 2012].

Инновации в научных исследований, проанализированные с помощью SAOM, уделяют внимание гендерной гомофилии [Lungeanu, Contractor, 2015], а также сетевым динамикам более молодых областей науки [Vermond и др., 2022], роли административных ресурсов университета на паттерны коллаборации [Roesler, Broekel, 2017], коллаборациям между университетом и индустрией [Chen и др., 2022] и диффузии инноваций [Liang, Liu, 2018].

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 коллаборации среди законодателей [Ingold, Fischer, 2014], использующие влияния рисков/ресурсов на принятие решений [Berardo, Scholz, 2010].

В поле исследований также входят статьи по анализу дипломатических связей между странами [Kinne, 2014]; международной кооперацией и работой интернациональных союзов в связи с проблемами координации [Kinne, 2013] и предсказания будущего Европы как структуры транснациональной сети [Thurner, Binder, 2009].

Лонгитюдный подход к сетям также полезен при анализе распространения правил и законов через институты, регионы и страны (например, законы вокруг международной торговли [Mohrenberg, 2017], а также коэволюции доступа к диджитал-инструментам, демократии и торговых связей [Rhue, Sundararajan, 2014]. Наконец, SOAM может быть применен для анализа политического действия [Redhead, Von Rueden, 2021] и исторического моделинга событий [Box-Steffensmeier, Jones, 2004].

2.2 Ограничения

SAOM обладает большими теоретическими основаниями, чем TERGM, и модель сочетает как социологические, так и статистические методы [Leifeld, Cranmer, 2019]. В то же время теоретическая укорененность модели может выступать и ограничением. Для начала, идея о том, что любые изменения в сети акторы выполняют только последовательно, а не одновременно не позволяет использовать данные из имейл сообщений, электрических систем, очень больших сетей и других ситуаций, когда “социологическая рамка не соответствует реальности” или она не может быть полностью фальсифицирована [Leifeld, Cranmer, 2022]. 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 задачами [Ceoldo, Snijders, Wit, 2023]. При этом, нельзя отрицать, что хотя процесс и ускоряется, работа с моделью требует больших временных затрат [Snijders, 1996], а получение необходимых данных высокого качества более финансово требовательно [Snijders, Van de Bunt, Steglich, 2010].

2.3 Перспективы

Перед исследователями стоят еще много методологических проблем, в том числе работа в модели с коррелирующими рандомизированными эффектами [Ceoldo, Snijders, Wit, 2023], и разработкой моделей с необнаруженной гетеронормативностью между акторами, которые позволят применить подход к более крупным сетям [Snijders, 2017].

Наконец, мы считаем, что SAOMs могут быть применены к нескольким направлениям исследований российского общества. До сих пор только в образовательной сфере были созданы литературные обзоры, направленные на выявления факторов академического успеха и поведения учеников с помощью SAOM [Докука, 2021; Докука, Валеева, 2015]. Мы предлагаем несколько потенциальных направлений развития.

Библиометрический анализ российского научного сообщества, которым занимается одна из исследовательских групп ANR-Lab уже использует лонгитюдные данные [Kim, Maltseva, 2021; Matveeva, Ferligoj, 2020]. 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 коллабораций в законодательных инициативах и торговых международных договорах. В данном случае, наиболее сложным этапом работы был бы сбор и предобработка данных.** ## Оптимизация и оценка Для оценки SAOM используются различные статистические методы, наиболее распространенным из которых является метод моментов (MoM) или метод максимального правдоподобия (MLE).

Метод моментов (MoM) – это метод имитационного моделирования, используемый для оценки параметров в SAOM, подробно описанный Снайдерсом (T. A. Snijders, 2001). 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 MoM 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 MoM может быть вычислительно трудоемким, а также требует тщательной настройки алгоритмов оптимизации. Предпринимаются попытки повысить эффективность и расширить спектр использования MoM. Например, развитие обобщенного метода моментов (GMoM) позволяет обогатить оценку временными

данными, вводя в нее в качестве параметров статистику из различных временных моментов (Amati et al., n.d.). Однако этот метод, как предполагают авторы, не может стабильно превосходить традиционный MoM, 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 (Besag, 1986). 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 (T. A. Snijders, 2001). Далее он предполагает, что адаптация спецификации модели, например, подходы, основанные на соседстве, с ограничениями на возможные связи между соседями (Pattison & Robins, 2002), подходы, основанные на латентном пространстве [Nowicki & Snijders (2001)](Hoff et al., 2002)(Schweinberger & Snijders, 2003), 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 (Lospinoso & Snijders, 2019) предлагают в качестве решения этой проблемы вспомогательные статистики (например, характеристики триад, транзитивность), не включенные в модель в явном виде. Они моделируют расстояние Махаланобиса между вектором вспомогательной статистики и оценкой модели с помощью симуляций Монте-Карло, повторно используя их из вычислений MOM 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:

$$AIC = -2 * loglikelihood + 2 * numberofmodelparameters$$

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

$$BIC = -2 * loglikelihood + log(samplesize) * numberofmodelparameters*$$

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 несошедшихся моделей от интерпретации следует отказаться.

· 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”.

2.4 Вырождение

Другой проблемой, возникающей при оценке SAOM, является вырождение. В работах Штрауса (Strauss, 1986), Снайдерса (T. A. B. Snijders, n.d.) и Хэндкока (Handcock, n.d.) 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 мультимодальных сетях, где типичные алгоритмы, обновляющие отдельные связи или структурные элементы, имеют ничтожно малую вероятность перемещения между модальными областями (T. A. B. Snijders, n.d.).

Для решения проблемы вырождения SAOM исследователи обычно используют различные стратегии (Handcock, n.d.), такие как проверка робастности, сравнение различных инициализаций модели,

предоставление дополнительных данных для обучения.

В этих практиках также отдается предпочтение байесовскому фреймворку (Nowicki & Snijders, 2001). Помимо уменьшения вырождения модели, он облегчает распространение неопределенности параметров на окончательный вывод и позволяет учитывать предварительные знания экспертов, если они существуют (Handcock, n.d.). Кроме того, Лоспиносо и др. (Lospinoso & Snijders, 2019) 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 (T. A. B. Snijders, n.d.).

3 Глава 2

See readings here Batagelj & Mrvar (n.d.)

Вот цитирование (**larranaga2013?**) и без скобок (**larranaga2013?**) — и сноска[3]. А вот пример ссылки на часть текста sec. 3.1.

3.1 Графики

Как отмечает Aoki (2007, p. 131):

> Пример цитаты. *А это пример курсива.*

3.1.1 Название подсекции

Заключение

[3]Внутренняя сноска

3.2 Современные подходы в области статистического сетевого анализа и моделирования: модели SIENA, ERGM, tERGM

3.2.1 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 (**leeHierarchicalGeneralizedLinear1996?**), временные ряды (**boxsteffensmeierTimeSeriesAnalysis2014?**), пространственный анализ (**wangPoissonlognormalConditionalautoregressiveModel2013?**) и моделирование многомерных распределений с использованием копула-функций (**genestJoyCopulasBivariate1986?**). 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 позволят 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 (**priceLittleScienceBig1963?**) и работы Гарфильда (**garfieldCitationIndexingIts1979?**), социологи представили несколько теорий, касающихся научного сотрудничества. Анализ саентометрии основан на эффекте Мэтью (**priceLittleScienceBig1963?**) и теории структуры малого мира (**desolapoolContactsInfluence1978?**), 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.

3.2.2 Стохастический и детерминированный подходы к сетевому анализу

В области сетевого анализа стохастические методы представляют собой кульминационные 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. Рис. 3).

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:

Детерминистские подходы к анализу сетей

Figure 3: Детерминистские подходы к анализу сетей

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 гомофилия.*** Эти свойства изучают характер связей между узлами на основе общих характеристик или атрибутов. Ассортативность изучает склонность узлов со схожими характеристиками к соединению, а гомофилия – 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.

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 (**denooyExploratorySocialNetwork2005?**).

2. ***Острова.*** Остров – это максимальная подсеть вершин, связанных между собой, значение которых больше, чем ребер, ведущих к вершинам вне такой подсети (**denooyExploratorySocialNetwork2005?**). Другими словами, это плотно связанные друг с другом узлы, отражающие локально важные участки сети. Алгоритм для поиска островов доступен в программе Pajek.

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 [(**denooyExploratorySocialNetwork2005?**)].

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, гомофилию 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 4.

Классификация основных стохастических подходов к анализу динамических сетей

Figure 4: 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 (T. A. Snijders, 2001), 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 (T. A. Snijders et al., 2007). 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. Хандкоком (**krivitskySeparableModelDynamic2014?**), 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 (**stadtfeldRejoinderDyNAMsGrounds2017?**), 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 (Butts, 2008), 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 событий.

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 (Kronegger et al., 2012; **wangQuantitativeExplorationReasons2018?**; **akbaritabarItalianSociologistsCommunity2020?**; **smithUnderstandingCollaborationPatterns2023?**). 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.

3.2.3 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 (Skyler et al., 2021). Дана 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 - θ вектор веществennозначных 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; - θ 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 (Lusher et al., 2012):

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.

3.2.4 Спецификация ERGM

Можно выделить три вида процессов формирования связей в сетях. Как показано на Рисунке 5, 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) (Lusher et al., 2012).

Figure 5: 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 (**priceLittleScienceBig1963?**). 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 нисходящие 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 (результирующей/результирующих 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 (**morrisSpecificationExponentialFamilyRandom2008?**). Данные термины определяются с использованием формулы R, которая включает как сеть, так и сетевые статистики:

y ~ $<term1> + <term2> + ...,$

где y – объект сети, а $<term1>$ и $<term2>$ – предопределенные термины, выбранные из списка (**morrisSpecificationExponentialFamilyRandom2008?**).

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 $(i \to j)$ и $(j \to i)$.
- *Asymmetric dyads* – сетевая статистика (только для ненаправленных сетей), обозначающая количество пар акторов, для которых существует либо $(i \to j)$, либо $(j \to i)$*.

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 "k-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 наблюдаемой структуры сети.

3.2.5 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, реципроктность 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 графами.
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 (**lusher2013?**). 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 (**robins?**) 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.

3.2.6 Темпоральный ERGM (TERGM)

Как определено в работе Лейфельда и Кранмера, TERGM развивают идею, заложенную в ERGM (**leifeldTemporalExponentialRandom2018?**). 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 (**leifeldTemporalExponentialRandom2018?**).

Граф зависимости TERGM, формально определяющий зависимость одной диады от другой, может моделировать зависимость между моделируемыми переменными в нескольких различных временных точках (**leifeldTemporalExponentialRandom2018?**). 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) (**hannekeDiscreteTemporalModels2010?**). Эти методы оценки удобно реализованы в пакете *btergm*, специально разработанном для среды статистических вычислений *R* (**leifeldTemporalExponentialRandom2018?**).

3.2.7 Исследование структур в наукометрических 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 (**block2018?**).

3.2.7.1 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 (**krivitsky?**) (**krivitsky2013?**) (**krivitsky2013?**) , (**handcock2016?**).

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 (**okamoto2015?**). Аналогичная ситуация складывается с исследователями в области компьютерных наук (**al-ballaa2019?**), информационно-поисковой сферы (**zhang2016?**) или исследователями научных инноваций из региона на западе Китая (**hou2023?**). 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 (**chakraborty2020?**), изучение словенских научных сообществ на примере 4 наук (физики, математики, биотехнологии и социологии) (**kronegger2011?**), исследование сотрудничества британских исследователей по финансируемым проектам (**smith2023?**). 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 (**chakraborty2020?**).

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 (**fanelli?**).

3.2.7.2 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 Трэвис Витшелл 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 (**whetshell2023?**). 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.

3.2.8 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 наукометрических 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 наукометрии.

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)(Мальцева & Fiala, 2023). 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. 3.2.8. 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)

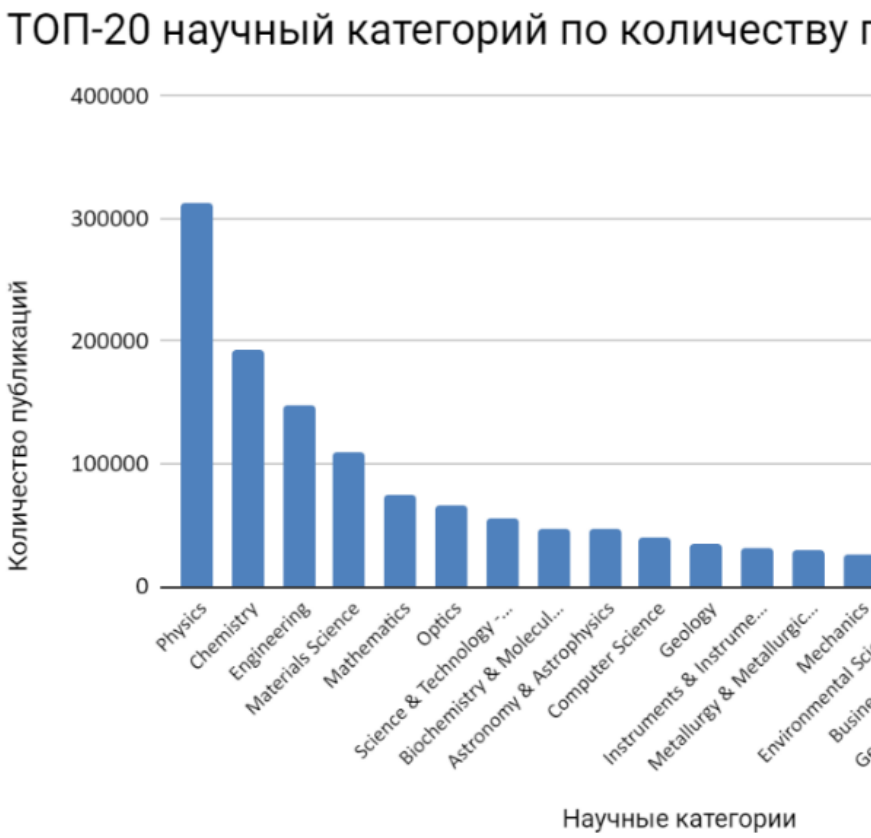
Ввиду отсутствия ограничений на этапе отбора данных и их большого размера, прежде чем приступить к анализу, мы были обязаны провести довольно крупный объем работ по предобработке данных. Сначала мы опишем общую структуру изначального массива данных, далее перейдем к процессу создания подсетов по отдельным 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. 6. 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 6: 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 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 мультиплицирует 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 «уникальных» ФИО авторов (в данных о российских социологах встречается до 8 вариаций ФИО одних и тех же авторов) являлось также минимальное число некорректных случаев соответсттвия 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 заменяли 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 стандартизированных 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 устранить 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%).

3.3 4.2.1 Библиометрический сетевой анализ коллабораций 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 (Shrum et al., 2007), 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 (Lundberg et al., 2006), данные для которых могут быть получены из наукометрических баз данных (Pike, 2010) (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 (Трофимова, 2023). 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 (Moed et al., 2018). 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 (Zhu & Liu, 2020). 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 (Pranckutė, 2021).

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 (eLibrary.ru, n.d.). 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. 7 представлена сеть цитирований с

источниками (изданиями), где публикую работы эти авторы. На первом месте с большим отрывом стоит журнал «Социологические исследования».

Сеть цитирований по источникам (изданиям) публикаций

Figure 7: Сеть цитирований по источникам (изданиям) публикаций

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 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 2: 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	Журнал	Публикации, шт.	Журнал	Цитирования, шт.	Журнал	Общая сила связей
3	Экономическая социология	540	International Journal of Intercultural Relations	557	Вестник Российского университета дружбы народов. Серия: Социология	47
4	Социологическое обозрение	517	Социологическое обозрение	396	Социология науки и технологий	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. Серия: Социология	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, шт.	Журнал	Цитирования, шт.	Журнал	Общая сила связей
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 8 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 8: 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 коллаборации, 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 ☒, как и каждый термин будет связан с другим весом ☒. 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 (Perianes-Rodriguez et al., 2016). 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 публиковавшиеся 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 3: Топ-10 представленных социологов

НПП	Автор	Публикации, шт.	Автор	Цитирования, шт.	Автор	Общая сила связей
1	Троцук И.В.	69	Инглахрт 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	Автор	Публикации, шт.	Автор	Цитирования, шт.	Автор	Общая сила связей
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 4: Топ-10 по публикационной продуктивности коллабораций

НПП	Организация	Публикации, шт.	Организация	Цитирования, шт.	Организация	Общая сила 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, шт.	Организация	Цитирования, шт.	Организация	Общая сила связей
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. 9). 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 коллабораций российских социологов за период 1992-2022 (фракционный счет, слева барьер – 5 и более публикаций, справа – 15 и более публикаций)

Figure 9: Сети соавторства коллабораций 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. 10). 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 отдалении 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 10: 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. 11). 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 11: 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 полицентричности 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 (Трофимова, 2023), при этом соотношение долей коллаборационной активности характеризуется перераспределением долей коллабораций – доля национальных коллабораций растет, в то время как доля международных снижается (Moed et al., 2018). 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.[4] 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 10 и 11, – 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.

[4] 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 коллаборативности.

Table 5: 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	**PUZANOVA_Z**	**16,53**	**33**
9	IVANOV_V	21,19	32
10	ROMANOVS_N	25,28	29
11	GORSHKOV_M	23,48	29
12	**GOLENKOV_Z**	**13,51**	**27**
13	BARSUKOV_S	22,33	25
14	LEVASHOV_V	21,21	25
15	TIKHONOV_N	21,17	25
16	**NARBUT_N**	**12,53**	**25**
17	FILIPPOV_A	20,00	22
18	SOKOLOV_M	18,33	22
19	TESLYA_A	21,00	21
20	STEPANOV_E	15,67	21

Коллаборации научных организаций (Рис. 12) 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 12: 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. Коллаборации по странам можно отнести к 2 категориям: двусторонние отношения (правая часть Рис. 3.3) и многосторонние коллаборации (левая часть Рис. 3.3), куда относятся как раз страны «традиционной» географии. Такие многосторонние коллаборации, безусловно, имеют больший потенциал охвата научного пространства, больше возможностей для привлечения новых участников коллабораций и более высокую публичность.

Сети коллабораций российских социологов по страновой принадлежности Анализ соавторства публикаций показал, что социологическое сообщество достаточно неоднородно, большое количество авторов не входит в ядро коллабораций. 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 коллаборации 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 коллабораций российских социологов является декомпозиция на «традиционную» и «новую» географии сотрудничества. Вторая особенность коллабораций – это количество стран-участников. С рядом стран выстраиваются только двусторонние коллаборации, а с другими российские социологи участвуют в многосторонних коллаборациях.

Анализ соавторства организаций продемонстрировал модель сотрудничества «ядро-периферия», где ядро представлено коллаборациями двух доминирующих организаций – Российской академии наук и Высшей школы экономики. Также данная модель сотрудничества характеризуется наличием группы традиционно представленных в коллаборациях 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 коллаборации российских социологов малочисленны и в основном представлены

российскими авторами с незначительным участием зарубежных ученых. Первым фактором, ограничивающим включенность в международные коллаборации, 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.

3.4 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 (Maltseva & Batagelj, 2020). В библиометрических исследованиях анализ соприсутствия ключевых слов является весьма популярным самостоятельным подходом, часто определяемым как картирование структуры знаний по соответствующему научному направлению (Павлова, 2023).

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) – 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 (Van Eck & Waltman, 2010). 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 (списки 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 (Aria & Cuccurullo, 2017). 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)

НПП	Ключевое 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

НПП	Ключевое слово	Соприсутствие в наборе данных (количество раз)	Ключевое слово	Общая сила связей
7	culture	71	gender	97
8	trust	61	trust	82
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», «democracry», «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 relations	54	international relations	44

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.

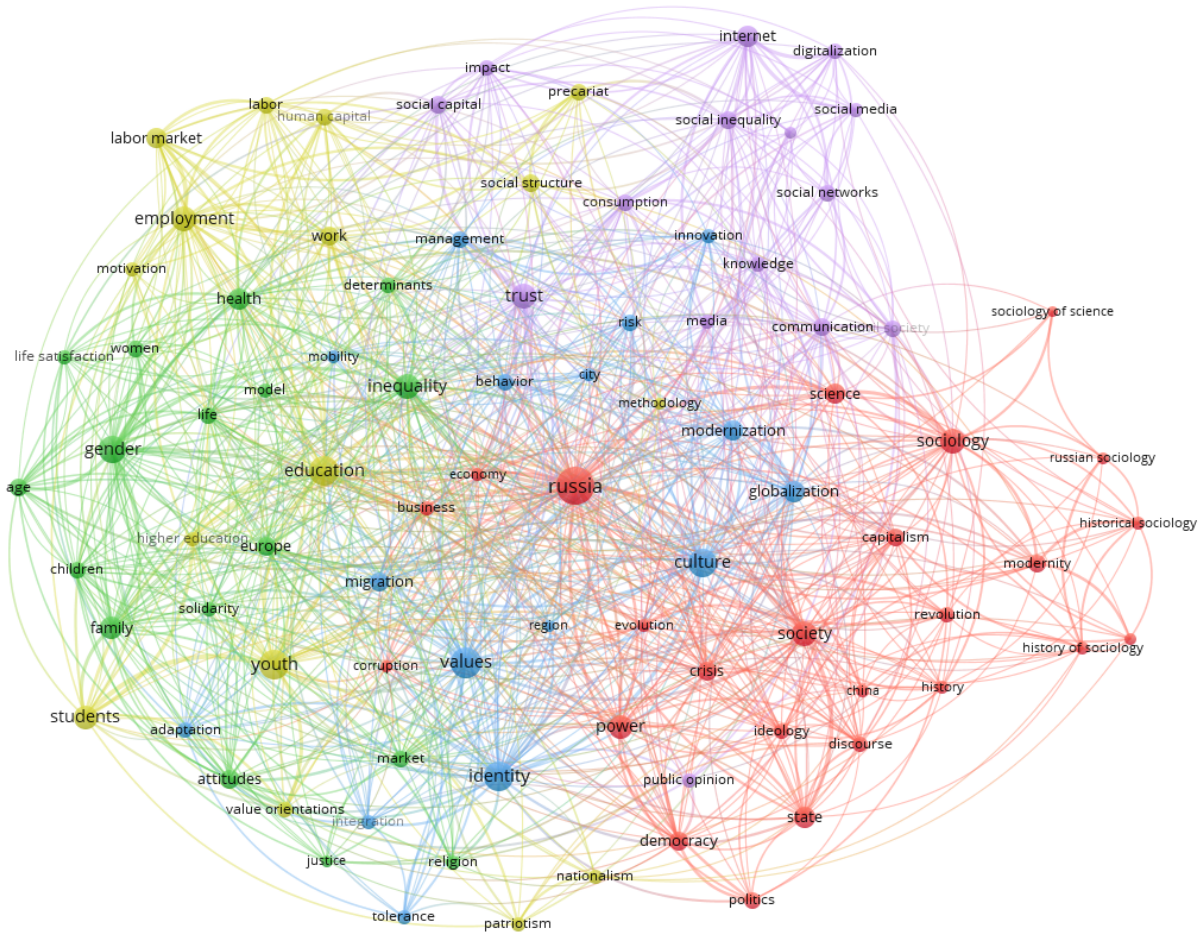


Рис. 1.1. Сеть соприсутствия ключевых слов (социология)

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.

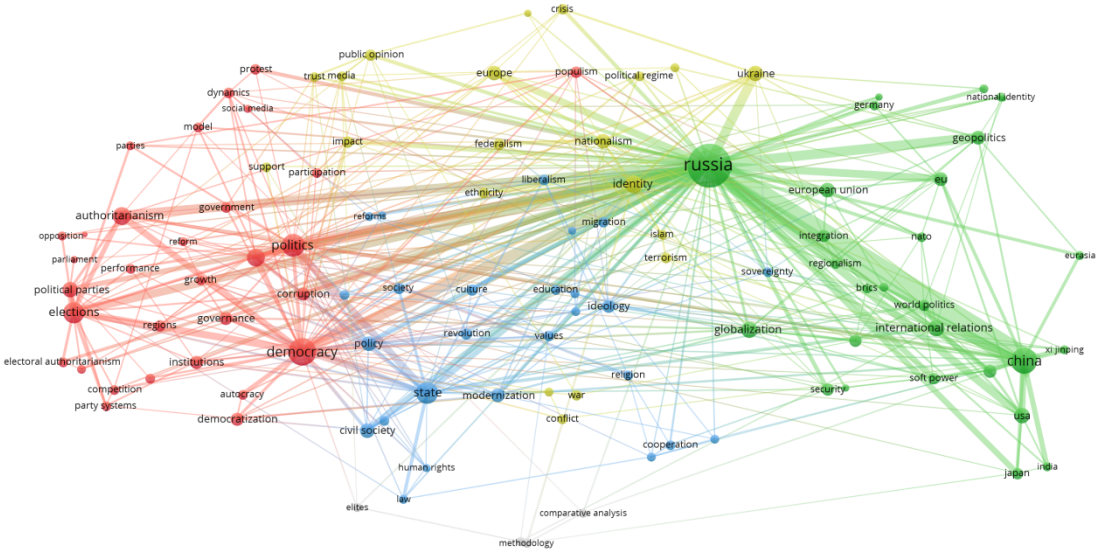


Рис. 1.2. Сеть соприсутствия ключевых слов (политология)

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 «укорененности» 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 года, являются долгоиграющими.

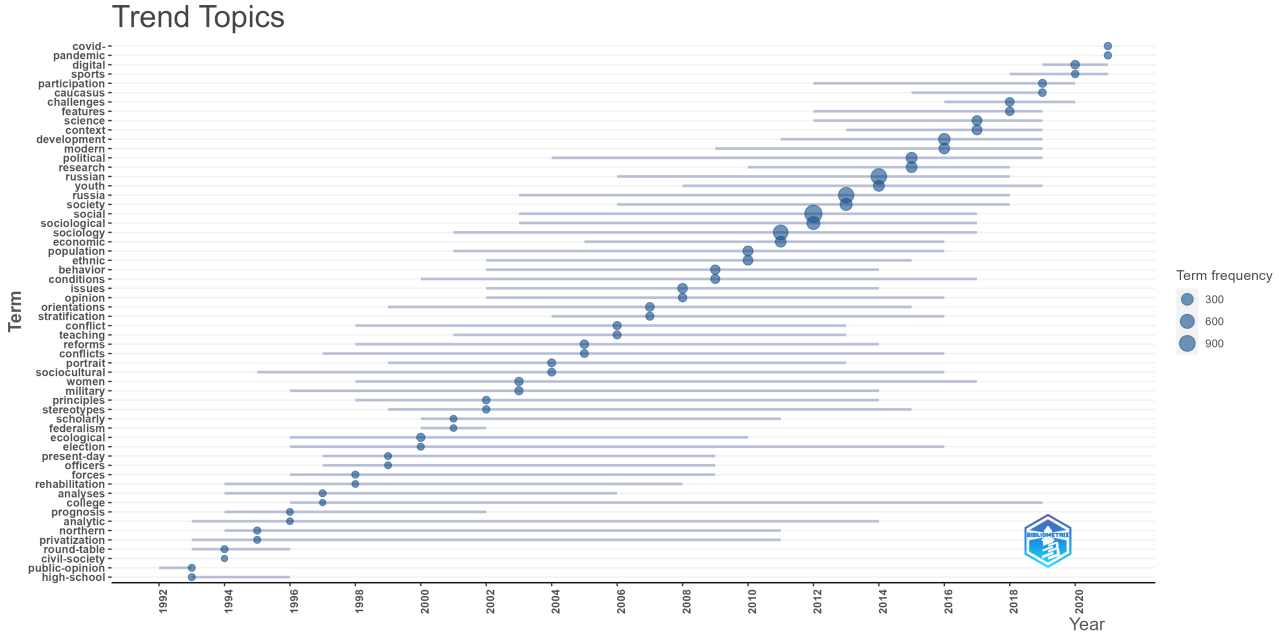


Рис. 2.1 Топ-2 популярных заголовка за каждый год (социология)

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 максимален 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.

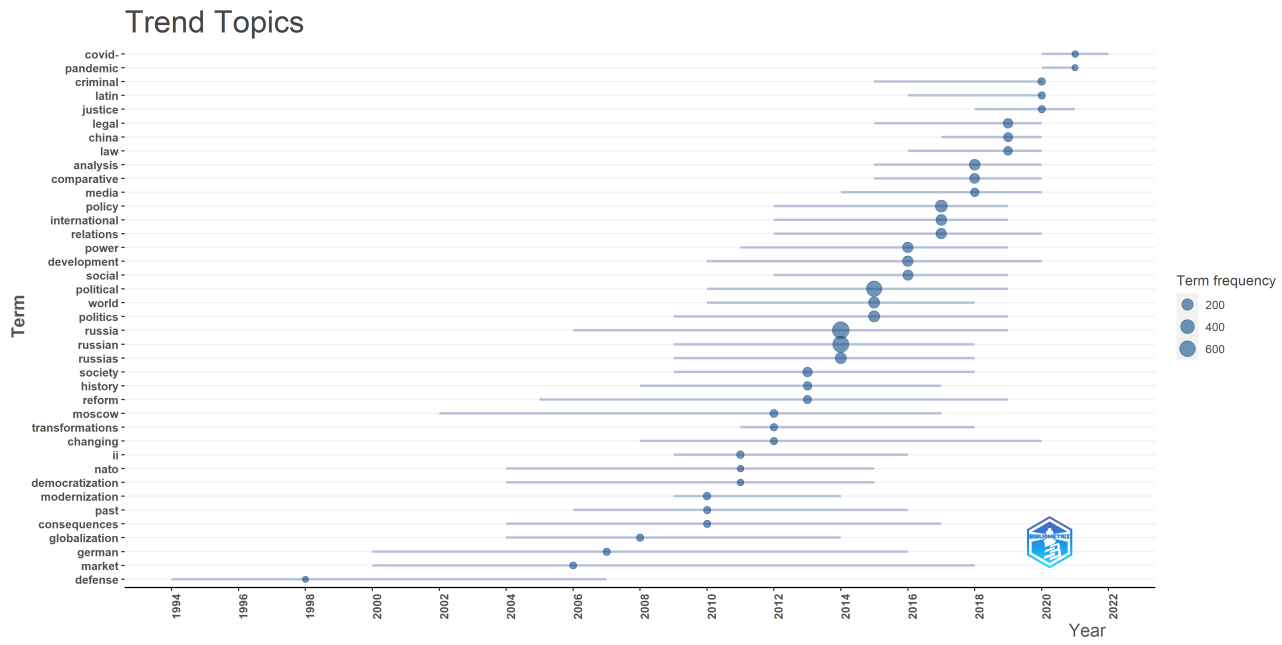


Рис. 2.2 Топ-3 популярных заголовка за каждый год (политология)

Выводы, полученные с помощью анализа частности упоминания тех или иных слов в заголовках публикаций, являются практичным 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 ограниченны, 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 (Cobo et al., 2011). 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.

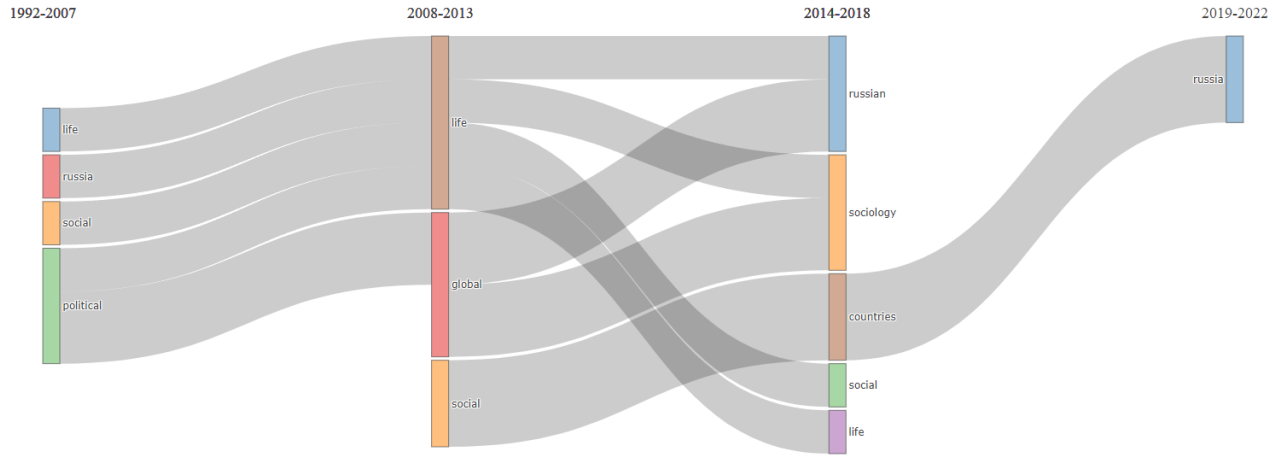


Рис. 3.1 Тематическая эволюция (социология)

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).

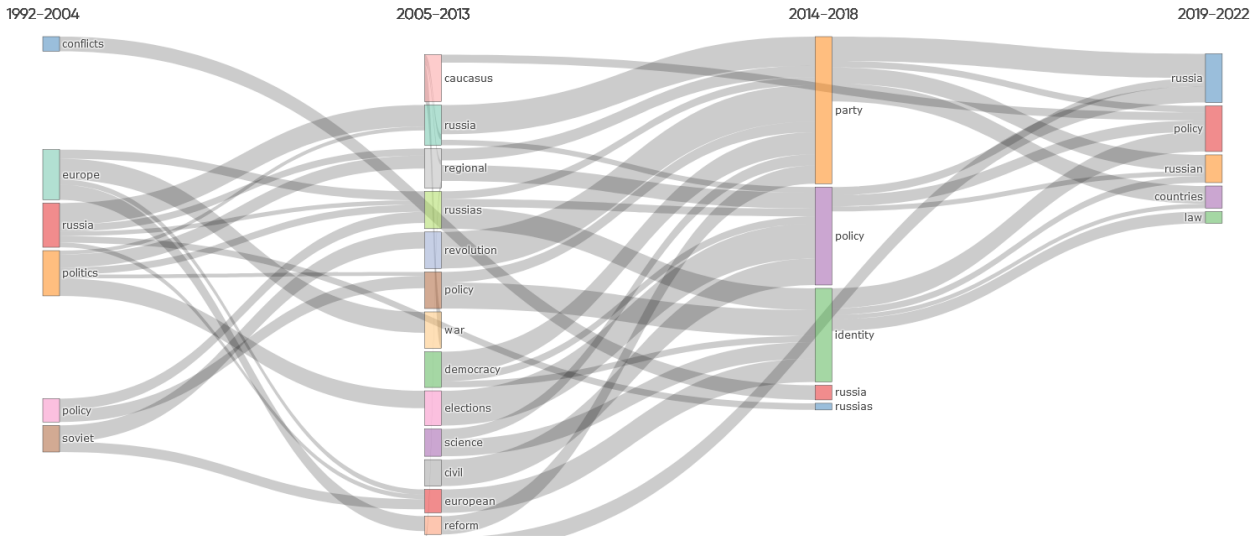


Рис. 3.2 Тематическая эволюция (политология)

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 сгенерированных 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 внутрироссийской 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 внутрироссийской политической трансформации. Глобальный сравнительный анализ сильно уступает как в плане проработанности, так и в плане релевантности для остального дисциплинарного контекста.

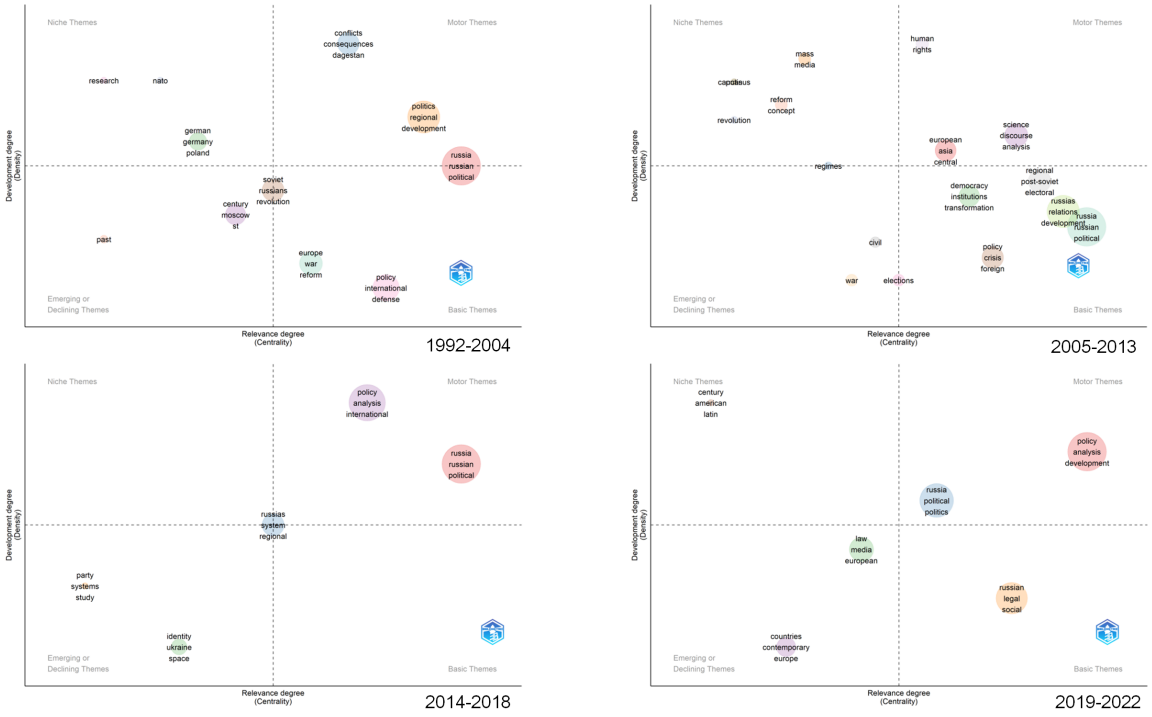


Рис. 4.2. Центральность-степень тематических кластеров во времени (политология)

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.

4 Честность и в стратегиях сплетен

4.1 ==Введение==

Данная статья (J. Wu et al., 2021) раскрывает интересный феномен сплетен без скобок J. Wu et al. (2021) — и сноска[5]. *А вот пример ссылки на часть текста* Заключение.

4.2 ==Сплетни==

Как отмечает J. Wu et al. (2021, p. 3):

> Пример цитаты. *А это пример курсива.*

4.2.1 Название подсекции

4.3 ==Заключение==

[5] J. Wu et al. (2021)

Figure 13: Картинка

Модульный шрифт Джозефа Альберса

Figure 14: Модульный шрифт Джозефа Альберса

[6]

Agrawal, P., Garg, V. K., & Narayanam, R. (2013). *Link Label Prediction in Signed Social Networks.* 2591–2597. https://doi.org/10.25781/KAUST-TZW78

Amati, V., Schönenberger, F., & Snijders, T. A. B. (n.d.). *Estimation of Stochastic actor-oriented models for the evolution of networksby generalized method of moments.*

Aoki, M. (2007). Endogenizing institutions and institutional changes*. *Journal of Institutional Economics*, *3*(1), 1–31. https://doi.org/10.1017/S1744137406000531

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959–975.

Batagelj, V., Ferligoj, A., & Doreian, P. (1999). *Generalized Blockmodeling.*

Batagelj, V., & Mrvar, A. (n.d.). *Pajek – Program for Large Network Analysis.*

Besag, J. (1986). On the Statistical Analysis of Dirty Pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, *48*(3), 259–302. JSTOR. http://www.jstor.org/stable/2345426

Butts, C. T. (2008). A relational event framework for social action. *Sociological Methodology*, *38*(1), 155–200.

Chen, X., He, Z., Chen, Y., Lu, Y., & Wang, J. (2019). Missing Traffic Data Imputation and Pattern Discovery with a Bayesian Augmented Tensor Factorization Model. *Transportation Research Part C: Emerging Technologies*, *104*, 66–77. https://doi.org/10.1016/J.TRC.2019.03.003

Cheng, C., Yang, H., King, I., & Lyu, M. (2012). *Fused Matrix Factorization with Geographical and Social Influence in Location-based Social Networks. 26*(1), 17–23. https://ojs.aaai.org/index.php/AAAI/article/view/8100

Cho, H., & Yu, Y. (2018). *Link prediction for Interdisciplinary Collaboration via Co-authorship Network. 8*(25). https://doi.org/10.1007/s13278-018-0501-6

Chuan, P. M., Son, L. H., Ali, M., Khang, T. D., Huong, L. T., & Dey, N. (2018). Link prediction in co-authorship Networks Based on Hybrid Content Similarity Metric. *Applied Intelligence*, *48*(8), 2470–2486. https://doi.org/10.1007/s10489-017-1086-x

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). 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*, *5*(1), 146–166. https://doi.org/10.1016/j.joi.2010.10.002

Coskun, M., & Koyutürk, M. (2015). *Link Prediction in Large Networks by Comparing the Global View of Nodes in the Network.* 485–492. https://doi.org/10.1109/ICDMW.2015.195

Daud, N. N., Ab Hamid, S. H., Saadoon, M., Sahran, F., & Anuar, N. B. (2020). *Applications of Link Prediction in Social Networks: A Review. 166*, 102716. https://doi.org/10.1016/j.jnca.2020.102716

eLibrary.ru. (n.d.). *Сравнение уровня публикаций российских ученых в базах данных Web of Science, Scopus и RSCI.* eLibrary.ru. Retrieved August 30, 2023, from https://elibrary.ru/wos_scopus_rsci.asp?

Farasat, A., Nikolaev, A., Srihari, S. N., & Blair, R. H. (2015). *Probabilistic graphical models in modern social network analysis. 5*(1), 1–18. https://doi.org/10.1007/s13278-015-0289-6

Getoor, L., Friedman, N., Koller, D., & Taskar, B. (2002). *Learning Probabilistic Models of Link Structure. 3*, 679–707. https://doi.org/10.1162/jmlr.2003.3.4-5.679

Gou, F., & Wu, J. (2021). *Triad Link Prediction Method Based on the Evolutionary Analysis with IoT in Oportunistic Social Networks. 181*, 143–155. https://doi.org/10.1016/j.comcom.2021.10.009

[6] J. Wu et al. (2021)

Handcock, M. S. (n.d.). *Assessing Degeneracy in Statistical Models of Social Networks.*

Hasan, M. A., & Zaki, M. J. (2011). A Survey of Link Prediction in Social Networks. In C. C. Aggarwal (Ed.), *Social Network Data Analytics* (pp. 243–275). Springer US. https://doi.org/10.1007/978-1-4419-8462-3_9

Haupt, M. R., Jinich-Diamant, A., Li, J., Nali, M., & Mackey, T. K. (2021). 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*, *21*, 100114. https://doi.org/10.1016/j.osnem.2020.100114

Hoff, P. D. (2009). Multiplicative Latent Factor Models for Description and Prediction of Social Networks. *Computational and Mathematical Organization Theory*, *15*(4), 261–272. https://doi.org/10.1007/s10588-008-9040-4

Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002). Latent Space Approaches to Social Network Analysis. *Journal of the American Statistical Association*, *97*(460), 1090–1098. https://doi.org/10.1198/016214502388618906

Huang, Z., Li, X., & Chen, H. (2005). *Link Prediction Approach to Collaborative Filtering.* 141–142. https://doi.org/10.1145/1065385.1065415

Koller, D., & Friedman, N. (2010). *Probabilistic Graphical Models: Principles and Techniques* (Nachdr.). MIT Press.

Kronegger, L., Mali, F., Ferligoj, A., & Doreian, P. (2012). Collaboration structures in Slovenian scientific communities. *Scientometrics*, *90*(2), 631–647. https://doi.org/10.1007/s11192-011-0493-8

Kuo, T.-T., Yan, R., Huang, Y.-Y., Kung, P.-H., & Lin, S.-D. (2013). Unsupervised Link Prediction using Aggregative Statistics on Heterogeneous Social Networks. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 775–783. https://doi.org/10.1145/2487575.2487614

Leguia, M. G., Levnajić, Z., Todorovski, L., & Ženko, B. (2019). Reconstructing Dynamical Networks via Feature Ranking. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *29*(9), 093107. https://doi.org/10.1063/1.5092170

Li, W.-J., Yeung, D. Y., & Zhang, Z. (2011). Generalized Latent Factor Models for Social Network Analysis. *International Joint Conference on Artificial Intelligence*, 1705. https://api.semanticscholar.org/CorpusID:6389369

Liben-Nowell, D., & Kleinberg, J. (2007). The link-Prediction Problem for Social Networks. *Journal of the American Society for Information Science and Technology*, *58*(7), 1019–1031. https://doi.org/10.1002/asi.20591

Liu, X., Kertkeidkachorn, N., Murata, T., Kim, K.-S., Leblay, J., & Lynden, S. (2018). Network Embedding Based on a Quasi-Local Similarity Measure. In X. Geng & B.-H. Kang (Eds.), *PRICAI 2018: Trends in Artificial Intelligence* (pp. 429–440). Springer International Publishing. https://doi.org/10.1007/978-3-319-97304-3_33

Liu, Y., Zhao, C., Wang, X., Huang, Q., Zhang, X., & Yi, D. (2016). The Degree-Related Clustering Coefficient and its Application to Link Prediction. *Physica A: Statistical Mechanics and Its Applications*, *454*, 24–33. https://doi.org/10.1016/j.physa.2016.02.014

Lospinoso, J., & Snijders, T. A. (2019). Goodness of fit for stochastic actor-oriented models. *Methodological Innovations*, *12*(3), 205979911988428. https://doi.org/10.1177/2059799119884282

Lü, L., & Zhou, T. (2011). Link Prediction in Complex Networks: A Survey. *Physica A: Statistical Mechanics*

and Its Applications, *390*(6), 1150–1170. https://doi.org/10.1016/j.physa.2010.11.027

Lundberg, J., Tomson, G., Lundkvist, I., Skar, J., & Brommels, M. (2006). Collaboration Uncovered: Exploring the Adequacy of Measuring University-Industry Collaboration through Co-authorship and Funding. *Scientometrics*, *69*(3), 575–589. https://doi.org/10.1007/s11192-006-0170-5

Lusher, D., Koskinen, J., & Robins, G. (Eds.). (2012). *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications* (1st ed.). Cambridge University Press. https://www.cambridge.org/core/product/identifier/9780511894701/type/book

Maltseva, D., & Batagelj, V. (2020). Towards a systematic description of the field using keywords analysis: main topics in social networks. *Scientometrics*, *123*(1), 357–382. https://doi.org/10.1007/s11192-020-03365-0

Marcot, B. G., & Penman, T. D. (2019). Advances in Bayesian network modelling: Integration of modelling technologies. *Environmental Modelling & Software*, *111*, 386–393. https://doi.org/10.1016/j.envsoft.2018.09.016

Moed, H. F., Markusova, V., & Akoev, M. (2018). Trends in Russian research output indexed in Scopus and Web of Science. *Scientometrics*, *116*(2), 1153–1180. https://doi.org/10.1007/s11192-018-2769-8

Mohan, A., Venkatesan, R., & Pramod, K. V. (2017). *A Scalable Method for Link Prediction in Large Real World Networks*. *109*, 89–101. https://www.sciencedirect.com/science/article/pii/S0743731517301600

Molokwu, B. C. (2021). *Social Network Analysis: A Machine Learning Approach*. University of Windsor, Canada.

Moradabadi, B., & Meybodi, M. R. (2018). *Link Prediction in Stochastic Social Networks: Learning Automata Approach*. *24*, 313–328. https://doi.org/10.1016/j.jocs.2017.08.007

Muniz, C. P., Goldschmidt, R., & Choren, R. (2018). Combining Contextual, Temporal and Topological Information for Unsupervised Link Prediction in Social Networks. *Knowledge-Based Systems*, *156*, 129–137. https://doi.org/10.1016/j.knosys.2018.05.027

Nasiri, E., Berahmand, K., Samei, Z., & Li, Y. (2022). Impact of Centrality Measures on the Common Neighbors in Link Prediction for Multiplex Networks. *Big Data*, *10*, 138–150. https://doi.org/10.1089/big.2021.0254

Nguyen, G. H., Lee, J. B., Rossi, R. A., Ahmed, N., Koh, E., & Kim, S. (2018). *Continuous-Time Dynamic Network Embeddings*. 969–976. https://doi.org/10.1145/3184558.3191526

Nowicki, K., & Snijders, T. A. B. (2001). Estimation and Prediction for Stochastic Blockstructures. *Journal of the American Statistical Association*, *96*(455), 1077–1087. https://doi.org/10.1198/016214501753208735

Özcan, A., & Öğüdücü, Ş. G. (2016). *Temporal Link Prediction Using Time Series of Quasi-Local Node Similarity Measures*. 381–386. https://doi.org/10.1109/ICMLA.2016.0068

Pattison, P., & Robins, G. (2002). 9. Neighborhood-Based Models for Social Networks. *Sociological Methodology*, *32*(1), 301–337. https://doi.org/10.1111/1467-9531.00119

Perianes-Rodriguez, A., Waltman, L., & Van Eck, N. J. (2016). Constructing Bibliometric Networks: A Comparison Between Full and Fractional Counting. *Journal of Informetrics*, *10*(4), 1178–1195. https://doi.org/10.1016/j.joi.2016.10.006

Pike, T. W. (2010). *Collaboration Networks and Scientific Impact among Behavioral Ecologists*. *21*(2), 431–435. https://doi.org/10.1093/beheco/arp194

Pranckutė, R. (2021). *Web of Science (WoS) and Scopus: The Titans of Bibliographic Information in Today's Academic World*. 1–59. https://doi.org/10.3390/publications9010012

Rettinger, A., Wermser, H., Huang, Y., & Tresp, V. (2012). *Context-Aware Tensor Decomposition for Relation*

Prediction in Social Networks. 2(4), 373–385. https://doi.org/10.1007/s13278-012-0069-5

Richardson, M., & Domingos, P. (2006). *Markov Logic Networks. 62*(1-2), 107–136. https://doi.org/10.1007/s10994-006-5833-1

Schweinberger, M., & Snijders, T. A. B. (2003). Settings in Social Networks: A Measurement Model. *Sociological Methodology*, *33*(1), 307–341. https://doi.org/10.1111/j.0081-1750.2003.00134.x

Shrum, W., Genuth, J., & Chompalov, I. (2007). *Structures of Scientific Collaboration.* The MIT Press.

Skyler, J. C., Bruce, A. D., & Jason, W. M. (2021). *Inferential Network Analysis* (Vol. 63). Cambridge University Press. https://doi.org/10.1007/s00362-022-01302-2

Snijders, T. A. (2001). The statistical evaluation of social network dynamics. *Sociological Methodology*, *31*(1), 361–395.

Snijders, T. A. B. (n.d.). *Markov Chain Monte Carlo Estimation of Exponential Random Graph Models.*

Snijders, T. A., Steglich, C., & Schweinberger, M. (2007). Modeling the co-evolution of networks and behavior. *Longitudinal Models in the Behavioral and Related Sciences*, *31*(4), 41–71.

Srilatha, P., & Manjula, R. (2016). *Paper Similarity Index based Link Prediction Algorithms in Social Networks: A Survey. 2*, 87–94.

Strauss, D. (1986). On a General Class of Models for Interaction. *SIAM Review*, *28*(4), 513–527. JSTOR. http://www.jstor.org/stable/2031102

Tan, Q., Liu, N., & Hu, X. (2019). Deep Representation Learning for Social Network Analysis. *Frontiers in Big Data*, *2*, 1–10. https://doi.org/10.3389/fdata.2019.00002

Van Eck, N., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538.

Wang, J., Ma, Y., & Yuan, Y. (2021). *Towards Fast Evaluation of Unsupervised Link Prediction by Random Sampling Unobserved Links* (No. arXiv:2002.09165). arXiv. https://doi.org/10.48550/arXiv.2002.09165

Wang, P., Xu, B., Wu, Y., & Zhou, X. (2014). *Link Prediction in Social Networks: the State-of-the-Art* (No. arXiv:1411.5118). arXiv. https://doi.org/0.1007/s11432-014-5237-y

Wang, S., Tang, J., Aggarwal, C. C., Chang, Y., & Liu, H. (2017). *Signed Network Embedding in Social Media.* 327–335. https://epubs.siam.org/doi/10.1137/1.9781611974973.37

Wu, J., Számadó, S., Barclay, P., Beersma, B., Dores Cruz, T. D., Iacono, S. L., Nieper, A. S., Peters, K., Przepiorka, W., Tiokhin, L., & Van Lange, P. A. M. (2021). Honesty and dishonesty in gossip strategies: a fitness interdependence analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1838), 20200300. https://doi.org/10.1098/rstb.2020.0300

Wu, Z., Lin, Y., Zhao, Y., & Yan, H. (2018). *Improving Local Clustering based Top-L Link Prediction Methods via Asymmetric Link Clustering Information. 492*, 1859–1874. https://doi.org/10.1016/j.physa.2017.11.103

Zhang, W., Liu, F., Xu, D., & Jiang, L. (2019). *Recommendation System in Social Networks with Topical Attention and Probabilistic Matrix Factorization. 14*(10), e0223967. https://doi.org/10.1371/journal.pone.0223967

Zhou, X., Liu, X., Wang, C., Zhai, D., Jiang, J., & Ji, X. (2021). *Learning with Noisy Labels via Sparse Regularization* (No. arXiv:2108.00192). arXiv. http://arxiv.org/abs/2108.00192

Zhu, J., & Liu, W. (2020). A Tale of Two Databases: the Use of Web of Science and Scopus in Academic Papers. *Scientometrics*, *123*(1), 321–335. https://doi.org/10.1007/s11192-020-03387-8

Мальцева, Д. В., & Fiala, D. (2023). Russian Publications in Web of Science: A Bibliometric Study.

(Препринт) COLLNET Journal of Scientometrics and Information Management.

Павлова, И. А. (2023). ПОСТРОЕНИЕ КАРТЫ СОПРИСУТСТВИЯ КЛЮЧЕВЫХ СЛОВ ПО ТЕМЕ «КАПИТАЛ ЗДОРОВЬЯ» В ПРОГРАММЕ VOSVIEWER. *Векторы Благополучия*, *49*(2), 38–54.

Трофимова, И. Н. (2023). *Международное сотрудничество российских исследователей: текущие позиции и тенденции: по данным Web of Science за 2018–2022 гг.* (No. 4). *32*(4), Article 4. https://doi.org/10.17323/1811-038X-2023-32-4-178-198