АННОТАЦИЯ

Выпускная квалификационная работа состоит из 4 глав, 37 с., включает 10 рис., 1 таблицу, 8 источников.

Ключевые слова: Стохастическое акторно-ориентированное моделирование, сетевой анализ, Stochastic Actor Oriented Models, социальные сети.

Объект исследования: Класс стохастических акторно-ориентированных моделей.

Цель работы: Построение и анализ стохастических акторно-ориентированных моделей на реальных данных и симуляция сетевой динамики.

Во введении отражена актуальность работы и поставлена проблема исследования стохастических акторно-ориентированных моделей. В первой главе представлен обзор литературы. Во второй главе рассмотрен класс стохастических акторно-ориентированных моделей в общем виде. В третьей главе описано применение САОМ к сети научного сообщества ТГУ. В четвертой главе рассмотрена имитационная модель сетевой динамики САОМ. Заключение содержит основные выводы и возможные направления дальнейшего развития данного исследования.

ОГЛАВЛЕНИЕ

ПЕРЕЧЕНЬ УСЛОВНЫХ ОБОЗНАЧЕНИЙ И ТЕРМИНОВ

Социальная сеть - объединение социальных акторов и их связей.

Актор - действующий субъект, совершающий действие, направленное на других акторов.

Ковариата - частная характеристика актора.

CAOM, SAOM – класс стохастических акторно ориентированных моделей (Stochastic Actor-Oriented Models).

Панельные данные, или лонгитюдные данные — используемые в социальных науках и эконометрике многомерные данные, получаемые серией измерений или наблюдениями за несколько периодов времени для одних и тех же компаний или людей.

ВВЕДЕНИЕ

В современном мире изучение социальных сетей приобретает всё большую значимость. Это связано с необходимостью анализа и оптимизации различных социальных процессов, будь то в науке, экономике или политике. С развитием сети интернет появилась возможность собирать и анализировать объективные данные о социальных сетях, и их структурах. Это вызвало развитие такой дисциплины, как сетевой анализ. В рамках сетевого анализа существует класс акторно-ориентированных моделей. Стохастическое акторно-ориентированное моделирование является мощным инструментом для анализа и предсказания сетевой динамики.

Объектом данного исследования является класс стохастических акторно-ориентированных моделей, а именно построение этих моделей, их применимость на реальных данных и построение имитационной модели. Предметом исследования выступают факторы, влияющие на динамику сети соавторства в Томском государственном университете.

Целью данной работы является построение и анализ стохастических акторно-ориентированных моделей на реальных данных и симуляция сетевой динамики. Для достижения поставленной цели можно выделить 2 глобальные задачи: Анализ реальных данных на примере сети научного соавторства Томского государственного университета и построение имитационной модели.

Подробнее про задачи данного исследования

1. Применение стохастической акторно-ориентированной модели (CAOM, SAOM) на реальных данных
    a. Выгрузка данных о авторах и их научных публикациях из *Tгу.Сотрудники*.
    b. Предварительная обработка данных. Выделение 3 снимков сети и векторов стажа, обработка некорректных данных и ошибок при загрузке.

c. Построение модели на основе социологических предположений и проверка гипотез
d. Анализ результатов

2. Построение имитационной модели САОМ
    a. Описание модели
    b. Разработка алгоритма
    c. Построение модели
    d. Анализ корректности работы модели

Изучение социальных сетей, особенно в контексте научного сотрудничества, позволяет понять ключевые механизмы, влияющие на формирование и эволюцию научных коллективов. В данном исследовании на примере сети соавторства в ТГУ будут изучены динамические процессы, влияющие на формирование научных связей. Использование стохастических акторно-ориентированных моделей позволяет не только описывать существующие структуры, но и предсказывать их будущее развитие, что является важным инструментом для управления научными коллективами и проектами. Написание же имитационной модели является начальным шагом для дальнейших исследований динамики разных социальных сетей, потому как это является необходимым этапом для проверки корректности при написании собственных реализация моделей САОМ, не описанных ранее.

1 Обзор литературы

В обзорной статье[1] рассматриваются стохастические акторно-ориентированные модели (САОМ) и дискретные экспоненциальные модели случайных графов (ERGM) для анализа динамики социальных сетей.

Эта работа представляет собой обзор двух статистических методов анализа данных, примененных для исследования социальной сети голландских школьников. Статья имеет особую значимость, так как является первой публикацией о САОМ на русском языке, что позволяет исследователям опираться на представленную авторами терминологию в рамках данного исследования.

В статье "Scientific collaboration dynamics in a national scientific system"[2] авторы исследуют взаимодействия исследователей в научной системе Словении в период с 1996 по 2010 годы. Работа охватывает широкий спектр вопросов, связанных с динамикой научного сотрудничества, применяя модели "малого мира" и эффект кумулятивного преимущества.

Модели «малого мира» (Small-World Models) характеризуются короткими средними путями между узлами сети, что позволяет моделировать эффективность распространения информации и взаимодействий в научном сообществе.

Модель Уоттса-Строгаца (Watts-Strogatz Model) является популярным примером модели «малого мира». Она позволяет генерировать граф, обладающий свойствами «малого мира», а также малой средней длиной кратчайшего пути и высоким коэффициентом кластеризации.

Эффект кумулятивного преимущества (Cumulative Advantage Process) описывает феномен, при котором более успешные и известные исследователи имеют тенденцию к получению большего числа новых коллабораций и ресурсов.

В статье формулируются социологические гипотезы о динамике сети научного сообщества и используются стохастические акторно-

ориентированные модели (СAOM) для анализа данных. Социологические гипотезы легли в основу части анализа реальных данных текущего исследования.

Анализ данных подтвердил многие из предложенных гипотез, показав, что структура научной сети в Словении действительно обладает свойствами "малого мира" и в ней проявляется кумулятивное преимущество. Использование СAOM позволило выявить важные закономерности в динамике научного сотрудничества и предложить рекомендации для улучшения управления научной деятельностью и поддержки исследователей.

«Stochastic Actor-Oriented Models for Network Dynamics»[3] является фундаментальной для данного исследования. Данная публикация представляет собой руководство по стохастическим акторно-ориентированным моделям, описывающее основные принципы, методы моделирования, построение моделей, их калибровку и интерпретацию результатов.

Работа Снейдерса и Пикапа является ключевым ресурсом для исследователей, занимающихся анализом динамики социальных сетей. Публикация не только описывает методологические аспекты СAOM, но и предоставляет практические рекомендации по их применению. Эти модели позволяют глубже понять механизмы формирования и развития социальных структур, что имеет важное значение для различных областей, включая социологию, психологию, экономику и информатику.

Руководство по пакету RSiena[4], подготовленное М. Р. Рутом, представляет собой детализированное методическое пособие для пользователей, работающих с программным обеспечением RSiena для анализа динамики социальных сетей. Данное руководство не только предоставляет практические инструкции по использованию RSiena, но и включает описание множества сетевых эффектов, а также интерпретацию результатов моделирования. Оно является незаменимым ресурсом для

исследователей, применяющих стохастические акторно-ориентированные модели (САОМ) для анализа сетей.

Работа Тома А. Б. Снайдерса по статистическому оцениванию динамики социальных сетей[5] является базовым текстом в области анализа социальных сетей (SNA). Он предоставляет всестороннюю схему для понимания того, как социальные сети эволюционируют со временем и вводит сложные статистические методы для моделирования и анализа этих изменений.

Снайдерс подчеркивает важность лонгитюдных данных, которые включают повторяющиеся наблюдения за одной и той же социальной сетью во времени. Этот подход позволяет исследователям улавливать динамику эволюции сети и понимать процессы, управляющие этими изменениями.

Значительный вклад Снайдерса состоит в разработке САОМ. Эти модели рассматривают изменения сети как результат решений и действий отдельных акторов, под влиянием их предпочтений и структуры сети.

В тексте детально описаны различные сетевые статистики (например, распределение степеней, меры центральности)

Таким образом, использование стохастических акторно-ориентированных моделей представляет собой перспективное направление в анализе социальных сетей. Применение SAOM к сети научного соавторства ТГУ позволит описать структуру и динамику сети, что важно для стратегического планирования и управления научными коллективами.

2 Теоретическая часть САОМ

2.1 Постановка задачи приводящей к САОМ

Рассмотрим некоторую социальную сеть. Изменение сети можно рассматривать как изменение актором своего «положения» в этой сети. Каждое изменение связей актора, а именно создание новой связи с другим актором, или разрушение существующей связи приводит к изменению всей сети.

Для сети дружбы между одногруппниками процесс создания или разрушения связи (в данном случае связь – считает ли актор другого актора своим другом) происходит в одностороннем порядке и не требует подтверждения принимающей стороны. То есть, можно считать своим другом того, кто не считает своим другом в ответ. Однако в случае сети научного соавторства необходимо подтверждение обеих сторон для создания связи. Также следует заметить, что в примере с сетью научного соавторства можно рассматривать факт связи как нерушимый, уже свершившийся факт, так и как процесс, в котором за какой-то период авторы писали вместе, а после перестали публиковаться.

Как при выборе того, с кем актор хочет изменить сеть, так и при выборе какое изменение внести (создание или разрушение связи), актор руководствуется некоторыми критериями оптимальности. Эти критерии могут зависеть как от внешних факторов, например возраста, пола, ученой степени, факта публикации в известном журнале, так и от сетевых факторов, например стремления создавать связи с самыми популярными акторами или увеличения вероятности создания связи между двумя акторами, если у них обоих уже есть связь с третьим актором.

При изучении реальных социальных сетей в большинстве случаев невозможно наблюдать за динамикой сети непрерывно. Пусть исследователю доступны лишь несколько снимков состояния сети. Эти снимки могут быть

сделаны в разные моменты времени, что позволяет анализировать изменения, произошедшие между ними.

Для анализа и предсказания динамики социальных сетей существует класс стохастических акторно-ориентированных моделей. Под понятием акторно-ориентированные подразумевается, что все изменения сети происходят из-за принятий решений акторами этой сети.

Использование SAOM позволяет исследователям моделировать и анализировать сложные социальные процессы происходящие в сети. Эти модели могут применяться для изучения различных типов социальных сетей, включая профессиональные, академические, дружеские и другие.

2.2 Описание САОМ

Модель изменения связи состоит из двух компонентов: времени и выбора. Время изменения определяется в терминах возможности изменения, а не факта изменения. Это означает, что акторам периодически предоставляется возможность пересмотреть свои связи, но они не обязаны их изменять. Если актор удовлетворен своим текущим состоянием, он не будет изменять свои связи.

Обозначим $X(t)$ как состояние сети в момент времени t. В любой момент времени в сети происходит не более одного изменения. Этими изменениями могут быть: создание или уничтожение связи.

Вероятность изменения состояния сети зависит от её текущего состояния, и не зависит от истории изменений. Таким образом $X(t)$ – непрерывная цепь Маркова. В этом контексте непрерывная цепь Маркова означает, что в каждый момент времени вероятность перехода сети в новое состояние определяется исключительно текущими связями и атрибутами акторов, без учета последовательности предыдущих изменений. На рисунке 1

можно увидеть пример изменения процесса $X(t)$ за один шаг. В данном случае изменением является создание новой связи между акторами 1 и 3. Где структура сети на рисунке 1 представлена на рисунке 2, новое ребро выделено кругом.

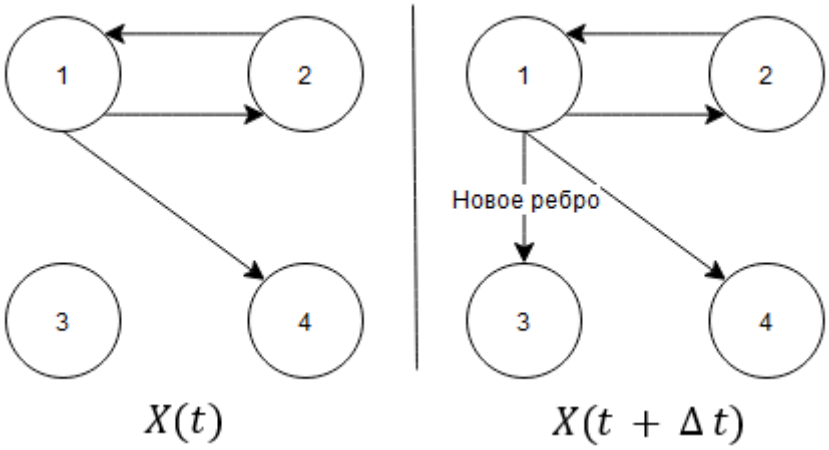

Рисунок 1 – Пример изменения процесса $X(t)$

	1	2	3	4
1	0	1	0	1
2	1	0	0	0
3	0	0	0	0
4	0	0	0	0

$X(t)$

	1	2	3	4
1	0	1	(1)	1
2	1	0	0	0
3	0	0	0	0
4	0	0	0	0

$X(t + \Delta t)$

Рисунок 2 – Пример представления процесса $X(t)$

Частота принятия решений актором i зависит от параметра λ_i . Обозначим $\lambda \triangleq \sum \lambda_i$.

$$P\{\text{Следущая возможность изменения наступит в } t + \Delta t \mid \text{Текущее время} - t\} = \exp(-\lambda_i * \Delta t)$$

Моменты времени распределены в соответствии с экспоненциальным распределением.

$$P\{\text{Следущая возможность изменения предоставлена актору } i\} = \frac{\lambda_i}{\lambda}$$

Эта формула соответствует модели «first past the post». Все участники имеют стохастическое время ожидания. Первый получивший возможность произвести изменение делает свой выбор и все начинается с начала, но уже в новом состоянии.

Совокупность выборов i актора образуют собой поток Пуассона δ_i с параметром λ_i.

Пример:

На рисунке 3 поток δ_1 , с интенсивностью λ_1

Рисунок 3 – Поток выборов актора 1

На рисунке 4 поток δ_2, с интенсивностью λ_2

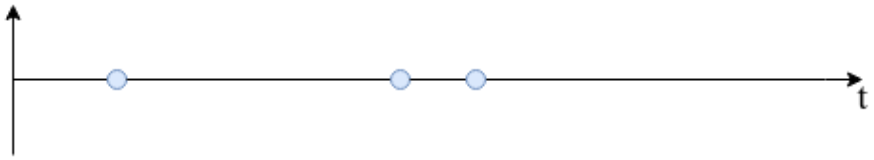

Рисунок 4 – Поток выборов актора 2

На рисунке 5 поток δ_3, с интенсивностью λ_3

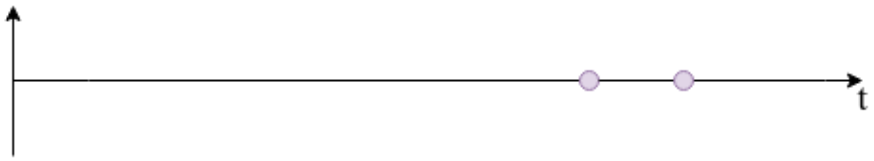

Рисунок 5 – Поток выборов актора 3

На рисунке 6 поток δ_n, с интенсивностью λ_n

Рисунок 6 – Поток выборов актора n

Образуют собой совокупный поток $N(t)$ с интенсивностью λ , представленный на рисунке 7.

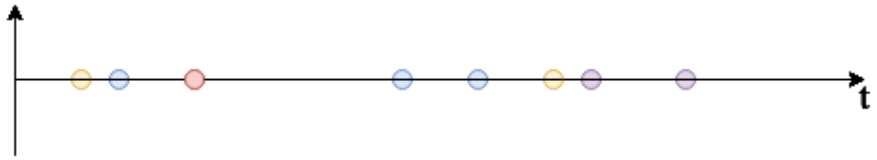

Рисунок 7 – Совокупный поток выборов всех акторов.

Сумма всех потоков – есть так же поток Пуассона с параметром λ, где времена между моментами скачков независимы и распределены экспоненциально.

Вероятность того, что i актор в выбранный промежуток времени примет решение о изменении сети:

$$P(X_{t+\Delta t} - X_t) = \lambda_i * \Delta t + o(\Delta t),$$

где $o(\Delta t)$ остаточный член высшего порядка.

Вероятность того, что какой то актор в выбранный промежуток времени совокупного потока $N(t_1)$, где $t \leq t_1 \leq t + \Delta t$ примет решение о изменении сети :

$$P(X_{t+\Delta t} - X_t) = \lambda * \Delta t + o(\Delta t).$$

У каждого актора есть n вариантов выбора, и всего $n - 1$ вариантов изменения сети. Построить\разрушить связь с кем то из других вершин, или ничего не делать. Для выбранного актора строится функция полезности для оценки вероятности выбора каждого из n вариантов. Функция полезности необходима для определения вероятности того, что при следующей смене связей данный актор перейдёт из состояния x в $x^{(\pm ij)}$.

Функция полезности:

$$f_i\left(x, x^{(\pm ij)}, \beta\right),$$

где i – индекс актора принимающего решение о связи с актором j,

x– текущее состояние сети, для графа и $i \neq j$,

$x^{(\pm ij)}$ – граф, который идентичен по всем рёбрам, кроме пары (i,j), для которой существование связи является переключателем $x^{(\pm ij)} = 1 - x_{ij}$, причём возможен случай когда $i = j$ это значит, что актор не изменит текущее состояние сети.

$f_i\left(x, x^{(\pm ij)}, \beta\right)$ определена на множестве всех пар сети $x^{(\pm ij)}$ и x. В интерпретации полезности функция полезности может рассматриваться как чистая полезность, которую получает актор i от перехода из состояния x в $x^{(\pm ij)}$.

Предположим, что в момент времени $t + \Delta t$, при текущей сети x у актора появилась возможность изменения в состояние $x^{(\pm ij)}$

$$\frac{exp\left(f_i\left(x, x^{(\pm ij)}, \beta\right)\right)}{\sum_h exp\left(f_i(x, x^{(\pm ih)}, \beta)\right)}.$$

Это распределение вероятностей выборов актора представляет собой модель множественного выбора при n вариантов решений у вершины. Подробнее про построение функции полезности:

Функция полезности:

$$f_i\left(x, x^{(\pm ij)}, \beta\right) = \sum_{k=1}^{K} \beta_k s_{ki}\left(x, x^{(\pm ij)}\right).$$

Функция полезности описывает полезность перехода из состояния x в $x^{(\pm ij)}$ для i актора.

В рассматриваемой модели функция полезности – есть линейная комбинация параметров β и функции сети, где функция сети описывает то,

как выглядит изменение сети из текущего состоянии в состояние новое с точки зрения актора i.

Функция сети:

$$s_{ki}\left(x, x^{(\pm ij)}\right),$$

где k – индекс параметра модели. Существует несколько реализаций функции сети [6] [3]. Некоторые из реализаций:

Базовый компонент исходящих степеней:

$$s_{1i} = \sum_j x_{ij}.$$

Он подобен постоянному члену в регрессионных моделях и всегда включается. Этот компонент балансирует между созданием и разрушением связей. Учитывая предыдущее состояние x, следующее состояние либо имеет на одну связь больше, либо на одну связь меньше, или они идентичны. Функция сети умножается на параметр β_1, он показывает вклад функции полезности в вероятность создания новой связи, вклад в разрушение существующей связи $-\beta_1$. Таким образом, роль эффекта степени исходящих связей в модели заключается в вкладе в пользу создания связей по сравнению с их разрушением. Обычно сети разрежены, поэтому возможностей для создания связей гораздо больше, чем для их прерывания.

Число взаимных связей:

$$s_{2i} = \sum_j x_{ij} x_{ji}.$$

Фундаментальный аспект почти всех направленных социальных сетей, потому что почти всегда существует некий обмен или другая взаимная зависимость.

Вероятность изменения связи $i \leftrightarrow j$:

$$P\{X(t+\Delta t) = x^{(\pm ij)} | X(t) = x\} = \frac{exp\left(f_i\left(x, x^{(\pm ij)}, \beta\right)\right)}{\sum_h exp\left(f_i(x, x^{(\pm ih)}, \beta)\right)},$$

где $X(t)$ – матрица смежности с элементами x_{ij} представляющая сеть в момент времени *t*.

Создание связи может быть как односторонней инициативой, так и совместной, в отличии от разрушения связи. Разрушение связи происходит в одностороннем порядке. Рассмотрим разные варианты моделирования возникновения связи.

$p_{ij}(x,\beta)$ – есть вероятность возникновения связи в разных сценариях.

1. Диктаторский сценарий
    a. Односторонняя инициатива:

$$p_{ij}(x,\beta) = \frac{exp\left(f_i\left(x,x^{(\pm ij)},\beta\right)\right)}{\sum_h exp\left(f_i\left(x,x^{(\pm ih)},\beta\right)\right)}$$

    b. Двусторонняя инициатива

$$p_{ij}(x,\beta) = \frac{exp\left(f_i\left(x,x^{(\pm ij)},\beta\right)\right)}{exp\left(f_i(x,x,\beta)\right)+exp\left(f_i\left(x,x^{(\pm ij)},\beta\right)\right)}$$

2. Взаимный сценарий:

$$p_{ij} = \frac{exp\left(f_i\left(x,x^{(\pm ij)},\beta\right)\right)}{\sum_h exp\left(f_i\left(x,x^{(\pm ih)},\beta\right)\right)}\left(\frac{exp\left(f_j\left(x,x^{(\pm ji)},\beta\right)\right)}{exp\left(f_j(x,x,\beta)\right)+exp\left(f_j\left(x,x^{(\pm ji)},\beta\right)\right)}\right)^{1-x_{ij}},$$

где $1-x_{ij}$ позволяет учесть факт существования или отсутствия связи в рассматриваемый момент времени.

$$\text{P\{ j примет предложение связи от i \}} =$$

$$= \frac{exp\left\{f_j\left(x,x^{(\pm ji)},\beta\right)\right\}}{exp\left\{f_j(x,x,\beta)\right\} + exp\left\{f_j(x,x^{(\pm ji)},\beta)\right\}}$$

Скорость изменения сети в случае направленной сети:

$$q_{ij} = \lambda_i p_{ij}(x,\beta).$$

Скорость изменения сети с односторонней инициативой, в случае ненаправленной сети:

$$q_{ij} = \lambda_i p_{ij}(x, \beta) + \lambda_j p_{ij}(x, \beta).$$

Скорость изменения сети с двусторонней инициативой, в случае ненаправленной сети:

$$q_{ij} = \lambda_{ij} p_{ij}(x, \beta) + \lambda_{ji} p_{ij}(x, \beta).$$

Q – матрица интенсивностей переходов цепи Маркова $X(t)$.

Таким образом Стохастические Акторно-ориентированные модели (САОМ) работают по следующему алгоритму:

1. Зададим $t = t_m, x = x(t_m)$.
2. Генерируем Δt в соответствии с экспоненциальным временем с параметром λ.
3. Если $t + \Delta t > t_{m+1}$ установить $t = t_{m+1}$ и остановиться.
4. Выбираем случайный $i \in 1..n$ используя распределение вероятностей $\frac{\lambda_i}{\lambda}$.
5. Выбираем случайным образом $x \in A_i(x)$, используя вероятности $p_i\left(x, x^{(\pm ij)}, \beta\right)$. $A_i(x)$ - множество всех состояний в которое может измениться сеть x в следствии решения актора i.
6. Задать $t = t + \Delta t$.
7. Задать $x = x^{(\pm ij)}$.
8. Вернуться к шагу (2).

На принятие решения может влияет множество эффектов, и параметров при этих эффектах. Дальнейшей задачей является оценка этих параметров.

2.3 Оценка параметров модели

Для оценки параметров используется имитационное моделирование. Реальная система заменяется моделью и производится множество повторений алгоритма. Когда модель в среднем будет хорошо согласовываться с данными, процесс оценки будет остановлен, и мы будем считать, что параметры оценены. Оценка параметров может быть произведена модернизированным методом моментов [4].

Обозначим оцениваемый параметр как

$$\theta = (\rho, \alpha, \beta).$$

Для каждого эффекта существует статистика, чувствительная к этому параметру. Для ρ_m, влияющего на общее количество изменений используется Расстояния Хэмминга,

$$D\big(X(t_{m+1}), X(t_m)\big) = \sum_{ij} X_{i,j}(t_m + 1) - X_{i,j}(t_m)$$

Для параметра α_k , обозначающего насколько сильно скорость изменения актора i зависит от $u_i, k(X)$, ковариат или позиционной характеристики актора, такой как исходящая степень $\sum_j x_{ij}$. Имеется в виду что скорость изменения может зависеть от пола, возраста, и тд.

$$A(X(t_{m+1}), X(t_m)) = \sum_{i,j} u_{i,k}\big(X(t_m)\big) + X_{i,j}(t_m + 1) - X_{i,j}(t_m).$$

Для функции полезности, где $s_{k,i}\big(x, x^{(\pm ij)}\big)$ не зависит от x, большие значения β_k будут приводить к увеличению вероятности перехода к сетям $x^{(\pm ij)}$ для которых значение $s_{k,i}\big(x, x^{(\pm ij)}\big)$ больше для любого актора i. Для оценки β_k статистика имеет вид

$$s_k\big(X(t_m)\big) = \sum_i s_{ki}\big(X(t_m)\big).$$

Комбинируя статистики и используя предположение о марковской цепи для наблюдаемых данных оценочные уравнения имеют вид.

$$D(X(t_{m+1}), X(t_m)) = E_\theta\{D(X(t_{m+1}), x(t_m) \vee X(t_m))\}, (m = 1, .., M-1),$$
$$\sum_{m=1}^{M-1} A_k, (k = 1, .., K_\alpha),$$
$$\sum_{m=1}^{M-1} s_k, \left(k = 1, .., K_\beta\right),$$

где K_α и K_β номера элементов α и β.

Для решения этих уравнений используется стохастическая оптимизация на основе алгоритма Роббинса-Монро. Алгоритм применяет многомерную версию алгоритма Роббинса-Монро с улучшениями, предложенными Поляком и Руппертом[10]. Техника "двойного усреднения" также применяется для улучшения результатов. Алгоритм реализован в пакете RSiena языка R.

Он состоит из трёх фаз:

1. Определение чувствительности ожидаемых статистик к параметрам.
2. Обновляет параметры с использованием симуляции динамики сети.
3. Используется для оценки приближения полученных уравнений и вычисления стандартных ошибок.

Для вычисления производных ожидаемых значений по отношению к параметрам используется метод функции оценки. Этот алгоритм является надёжным, но затратным по времени[10].

3 Применение модели на реальных данных

3.1 Объект исследования

Объектом исследования являются научные группы и команды Томского Государственного Университета, а конкретно данные, собранные о сотрудниках ТГУ, их научных публикациях и сетевых связях. Используя библиометрический анализ и стохастическое акторно-ориентированное моделирование, можно выявить факторы, влияющие на формирование научного взаимодействия внутри этих групп и команд.

Данные, собранные с веб-сайта «ТГУ.Сотрудники», включают в себя информацию о научном опыте сотрудников и список их публикаций, что позволяет построить сеть соавторства и исследовать взаимосвязи между учеными. Анализируя эти данные, мы можем выявить ключевые факторы, влияющие на формирование научного сообщества и его продуктивность.

3.2 Применимость САОМ в исследовании научных групп

Научное взаимодействие представляет собой процесс сотрудничества между учёными с целью достижения общей научной цели. Однако, как и любой другой процесс, научное взаимодействие может изменяться под влиянием различных факторов, которые могут способствовать или, наоборот, препятствовать его формированию и развитию.

Основными факторами, влияющими на формирование и разрушение научного взаимодействия, которые можно выделить в рамках взаимодействия в Томском Государственном Университете:

1. Общая научная тематика: Учёные, работающие в одной области науки, имеют больше возможностей для сотрудничества, чем те, кто занимаются разными научными направлениями.

2. Финансирование: Наличие достаточных финансовых ресурсов может способствовать формированию научных групп и команд. Однако, в случае недостатка финансирования, учёные могут ощутить трудности в формировании и поддержании научных контактов.
3. Личностные факторы: Личностные характеристики учёных, такие как общительность, коммуникабельность и т.д., могут оказывать существенное влияние на формирование и поддержание научных связей.
4. Технологический прогресс: Технологический прогресс может предоставить новые возможности для научного сотрудничества.
5. Наличие научной инфраструктуры: Наличие необходимой инфраструктуры, такой как лаборатории, библиотеки, научные журналы, тоже является важным фактором для формирования научных групп и команд.

САОМ - это метод социально-сетевого анализа, который позволяет моделировать и предсказывать эволюцию социальных сетей на основе поведения и взаимодействий индивидуальных акторов в них. САОМ используется для изучения процессов формирования социальных связей в группах и командах, а также для оценки влияния различных факторов на эти процессы.

Применительно к исследованию научных групп и команд, САОМ может быть использован для моделирования процессов формирования научных связей между учёными в рамках группы или команды.

Используя САОМ возможно выявить такие факторы, влияющие на формирование связей, как: влияние стажа на то, с какой частотой актор будет принимать решения относительно изменения сети, с какой частотой изменяется сеть, как влияет научное подразделение актора на динамику сети,

так и влияние сетевых характеристик на изменение рассматриваемой сети и т.д.

Преимущества CAOM включают возможность моделирования процессов на уровне индивидуальных акторов, учёт динамической природы социальных связей, а также возможность оценки влияния различных факторов на эти процессы. Однако CAOM также имеет некоторые ограничения, такие как необходимость большой вычислительной мощности для моделирования и ограниченная возможность предсказания долгосрочных эффектов.

Таким образом, CAOM является мощным инструментом для исследования научных групп и команд, который может помочь в выявлении факторов, влияющих на формирование научных связей и определении эффективных мер по их сохранению а так же оптимизации научного взаимодействия.

Алгоритмическая сложность моделей CAOM не позволяет просто «перебрать» возможные эффекты, и варианты моделирования принятия решения (сценарий, по которому задается вероятность изменения ребра в момент принятия решения). Исследование с использованием CAOM предполагает наличие гипотезы о виде социальной сети, её структуре и факторах, влияющих на динамику рассматриваемой сети.

3.3 Формирование гипотез о динамике рассматриваемой сети

Формирование начальных гипотез, а так же формирование социологических выводов – это задача прикладных специалистов. Будут использованы некоторые предположения выдвинутые в подобном исследовании для национальной научной системы Словении [2].

Согласно исследованию Anuška Ferligoj, Luka Kronegger, Franc Mali, Tom A B Snijders, Patrick Doreian в журнале Scientometrics (2015, №104, с. 989–990):«В целом, мы предлагаем следующие механизмы влияния на научное соавторство:

- встраиваемость в сеть: соавторы соавторов стремятся стать соавторами соавторов;
- преимущественная привязанность: авторы предпочтительно ищут соавторов, у которых уже много соавторов;
- институциональная встроенность (принадлежность к одной исследовательской группе и одной научной дисциплине, возрастное сходство также может подпадать под эту категорию, поскольку означает принадлежность к общей когорте учёных, которые взаимодействуют друг с другом больше, чем представители разных когорт) и контрольные переменные, в частности стаж. »

3.4 Описание модели

Необходимо учесть факт, что сеть соавторства является ненаправленной:

$$X_{i,j} = X_{j,i},$$

где $X_{i,j} \in 0{,}1$ является элементом графа и обозначает связь $i \leftrightarrow j$.

В силу ненаправленности сети формирование связи между акторами происходит по одному из двух сценариев:

1. Односторонняя инициатива: Выбирается один актор i, который получает возможность произвести изменение.

2. Двусторонняя инициатива: Выбирается упорядоченная пара акторов (i, j), и получает возможность принять новое решение о связи $.i \leftrightarrow j$.

Процесс выбора моделируется по одному из двух сценариев:

D. Диктаторский(Dictatorial): Один актор может навязывать решение о связи другому актору.

M. Взаимный(Mutal): Оба актора дают согласие на существование связи между ними.

Согласно выдвинутым социологическим предположениям[2] формирование связи будет происходить согласно двусторонней инициативы, и оба актора будут давать согласие на создание связи.

Факторы влияющие на оценку полезности создания связи:

1. Влияние количества исходящих связей
2. Влияние транзитивных троек
3. Влияние стажа на эффект транзитивных троек
4. Влияние количества исходящих связей на частоту принятия решений.

3.5 Оценка сформированных гипотез

3.5.1 Оцененные параметры при эффектах

Коэффициент сходимости модели достаточно низок 0.0508, для того, чтобы считать модель сходящейся[4]. Что означает, что выдвинутые предположения о виде модели корректно описывают представленные снимки сети. Так же абсолютное значение *Convergence t-ratio*[13] для каждого параметра меньше 0,1, что говорит о значимости каждого из эффектов. В таблице 1 представлены оценки, стандартные отклонения и Convergence t-ratio для каждого эффекта.

Таблица 1 - Оценённые значения эффектов

Эффект	Оценка	Стандартное отклонение	Convergence t-ratio
Интенсивность в 1 периоде	0.2317	0.0053	
Интенсивность в 2 периоде	0.1678	0.0091	
Влияние количества исходящих связей	-3.2147	0.0224	0.0157
Влияние транзитивных троек	0.4268	0.0140	-0.0005
Влияние стажа на эффект транзитивных троек	-0.0018	0.0006	0.0112
Влияние количества исходящих связей на частоту принятия решений	1.6070	0.0224	0.0157

3.5.2 Интерпретация полученных оценок параметров

Интерпретация для эффектов, использующих только сетевые данные, и эффектов, использующих ковариаты, различается.

Необходимо учитывать, что функция оценки полезности принятия решения при использовании ковариаты задана формулой:

$$f_i^{beh}(x,z) = \sum \beta_k^{beh} s_{ik}^{beh}(x,z),$$

$$s_{1k}^{beh}(x,z) = z_i s_{ik}^{0}(x,z),$$

где $s_{ik}^{0}(x,z)$ не зависит от z_i причём может зависеть от z_j для другого актора j. Она отражает полезность для i актора. Используется для сравнения полезности различных изменений.

Рассматривая же функцию полезности для сетевых структур $f_i^{net}(x) = = \sum \beta_k^{net} s_{ik}^{net}(x)$, когда i актор делает изменение из x в $x^{(\pm ij)}$, и $x_a x_b$ два возможных результата шага то отношение вероятностей этих вариантов есть:

$exp\left(f_i^{net}(x_a) - f_i^{net}(x_b)\right)$. Следовательно сетевые параметры стоит расценивать как логарифмические.

Для оценки зависимых переменных (либо X_{ij}, либо s_{ik}) необходимо учитывать, что возможные изменения в переменной $+1,0,-1$, и оценки не должны выходить за пределы допустимого диапазона. Параметры целевой функции это вклады в логарифмические вероятности увеличения зависимой переменной на 1 единицу при увеличении эффекта на 1 единицу.

Пример:

Рассчитан эффект *среднее число связей того, к кому направлена связь* он равен $1{,}1561$. Это означает, что при сравнении двух акторов равных во всех отношениях, кроме того, что связи первого в среднем на 1 выше по шкале рассматриваемой ковариаты (Индекс Хирша, и тд), чем у второго, шанс увеличения значения ковариаты по сравнению с отсутствием изменений (в рамках одного шага относительно ковариаты) выше в $exp(1{,}1561) = 3{,}17751676428$ раза, чем для второго.

Не любой эффект можно трактовать как в приведённом примере.

Пример:

Для эффекта *влияние количества исходящих связей* для актора i x_{i+}. Пусть число связей, чьи значения x_j меньше, равны или больше значения z_i для i, обозначим через a, b и c. Обозначим диапазон (максимальное минус минимальное значение) через r . Тогда, в случае минимального шага в отношении поведения, вклады суммарного эффекта сходства в лог-вероятности изменений $+1,0,-1$ определяются следующим образом $\beta_k^{beh}(a-b-c)/r$ и $\beta_k^{beh}(c-a-b)/r$. Вклады для среднего эффекта сходства составляют $\beta_k^{beh}(a-b-c)/\mathrm{r}$ и $\beta_k^{beh}(c-a-b)/\mathrm{r}$. Это показывает, что влияние связей на эффекты сходства зависит только от того, имеют ли они большие или меньшие значения, чем у рассматриваемого актора, а не от того, насколько эти значения больше. Это также показывает, что для эффектов сходства важна дисперсия ценностей связей, а не

а не только их среднее значение, в то время как для эффекта среднего изменения имеет значение только среднее значение.

Не все эффекты могут быть описаны как изменение некоторой функции полезности. В таких случаях используются элементарные эффекты [5].

Элементарный эффект - это вклад в создание или поддержание связи, определённый напрямую, то есть без выражения его на основе изменения какой-либо функции оценки. Это означает, что элементарные эффекты более общие, чем эффекты оценки, и все эффекты могут быть представлены как элементарные эффекты. Однако для удобства интерпретации предпочтительно использовать формулировку функции оценки.

Элементарные эффекты могут применяться аналогично к созданию и поддержанию связи; или они могут применяться исключительно к созданию связи или исключительно к поддержанию связи.

Пример ситуации, когда невозможно описать как изменение функции полезности - тенденция к созданию связных троек (транзитивные замыкания). Если актор i является центральным в тройке, то связи, приводящие к замыканию, представлены как $i \to j$ и $i \to h$. Однако первая связь подразумевает создание пути $i \to h \to j$, в то время как вторая связь означает установление связи с актором h. Актор h, в свою очередь, делает тот же исходящий выбор для третьего актора j, что указывает на структурную эквивалентность; таким образом, это различные процессы.

Таким образом полученные оценки можно интерпретировать так:

Влияние количества исходящих связей указывает на то, что с увеличением исходящих степеней снижается вероятность появления новой связи.

Зависимость скорости (параметра λ_i для актора i) *от количества исходящих степеней*: с увеличением числа связей актор увеличивает частоту принятия решений.

Транзитивные тройки: показывает количество транзитивных паттернов относительно актора i. Для двух акторов, где каждый из них является соавтором с i индивидуально, вероятность стать соавторами увеличивается в $exp(0.4268) = 1.5324$ раза.

Влияние стажа на эффект транзитивных троек означает, что с увеличением стажа ослабевает эффект транзитивных троек.

4 Построение имитационной модели

4.1 Описание модели

Исследование реальных данных проводилось с использованием пакета RSiena. RSiena предоставляет широкий спектр различных реализаций компонентов модели и распространяется с открытым исходным кодом, а также имеет активную поддержку сообщества для написания пользовательских компонентов.

Пользовательский компонент может представлять собой эффект, например, влияние временного ряда научных конференций на степень кластеризации всей сети. Такой компонент должен включать в себя анализ временного ряда, использование dummy variable и классический компонент из RSiena - связные тройки.

Для проверки корректности пользовательских эффектов необходима имитационная модель САОМ.

Была построена модель для направленной сети, где возможные результаты решения актора включают создание новой связи, разрушение существующей связи или оставление сети в том виде, в котором она есть в текущий момент времени. Разрушение связи происходит в одностороннем порядке, а создание связи также осуществляется в одностороннем порядке.

Вероятность изменения связи $i \leftrightarrow j$:

$$P\{X(t+\Delta t) = x^{(\pm ij)} | X(t) = x\} = \frac{exp\left(f_i\left(x, x^{(\pm ij)}, \beta\right)\right)}{\sum_h exp\left(f_i(x, x^{(\pm ih)}, \beta)\right)},$$

Функция полезности – есть линейная комбинация функций сети и их параметров. Для имитационной модели были взяты такие сетевые эффекты как:

1. Влияние количества исходящих связей,
2. Влияние количества общих связей.

В таком случае функция полезности данной модели имеет вид

$$f_i\left(x, x^{(\pm ij)}, \beta\right) = \beta_2 * \left(\sum_j x_{ij}\, x_{ji}\right) + \beta_1 * \left(\sum_j x_{ij}\right) \quad (1)$$

4.2 Допущения

Из-за отсутствия использования в имитационной модели эффектов, связанных с ковариатами, выбор актора, принимающего решение, можно упростить. Для каждого актора интенсивности рассматриваются одинаковыми, что позволяет применить равномерное распределение вероятностей при выборе актора на этапе принятия решения.

Сеть представлена квадратной матрицей, заполненной нулями и единицами. Начальное состояние сети на момент t_0 генерируется случайно с использованием равномерного распределения.

Интервалы времени между снимками сети одинаковы по длительности.

4.3 Описание работы имитационной модели

В начале генерируется начальное состояние сети, представленное матрицей смежности размерности N , где N – количество акторов в сети. Начальная матрица заполняется единицами с заданным пользователем коэффициентом разреженности. Затем стимулируется социальная динамика до момента времени начала следующего снимка сети t_{m+1}.

Процесс повторяется дважды, используя каждый раз снимок, полученный на предыдущем шаге, таким образом, в результате работы получается три снимка сети.

Подробнее про симуляцию сетевой динамики. Положим $t_m = 0, t_{m+1}$ задаётся пользователем, вектор параметров β при эффектах задаётся пользователем. Генерируем Δt экспоненциально, с параметром λ . Пока $t_m + \Delta t$ меньше t_{m+1} используя распределение акторов выбирается актор принимающий решение о изменении сети в момент $t_m + \Delta t$. Если $t_m + \Delta t$ меньше t_{m+1} то генерируются все возможные состояния сети, в которые

можно перейти из текущего состояния. Выбранный актор (актор i) используя распределение вероятностей переходов во все возможные варианты новой сети $x^{(\pm ij)}$ принимает решение о изменении сети. Процесс заканчивается когда $t_m + \Delta t > t_{m+1}$.

Для симуляции принятия решения строится вектор полезностей для каждого актора при зафиксированном ii акторе. Каждый элемент этого вектора отражает полезность изменения сети из текущего значения в следующее (минишаг). Считается функция полезности для каждого возможного изменения, где конкретный вид функции полезности задан формулой (1).

4.4 Оценка работы имитационной модели

Для оценки корректности работы имитационной модели необходимо передать полученные снимки сети в пакет RSiena, описав модель. Далее коэффициент сходимости должен быть ниже $0{,}1$[4], абсолютное значение Convergence t-ratio[4] для каждого параметра меньше $0{,}25$, а так же 2 среднеквадратических отклонения не должны превышать значение оценки параметра.

Был построен алгоритм, запускающий имитационную модель, передающий результат работы (снимки сети) в пакет RSiena и сохраняющий результат оценки параметров, СКО каждого параметра, коэффициент сходимости модели, и Convergence t-ratio. На рисунке 8 можно увидеть как возрастает количество связей (голубых ребер) между акторами (красными вершинами) в результате симуляции сетевой динамики.

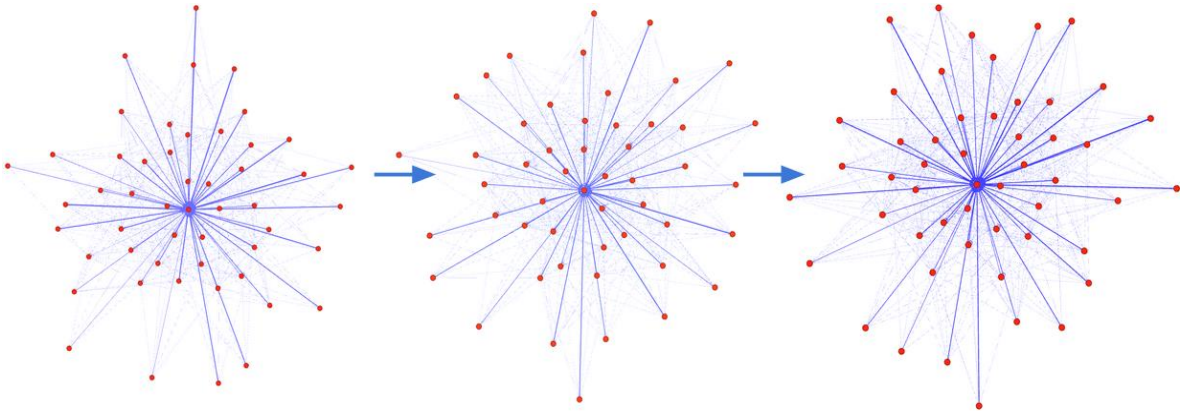

Рисунок 8 – Визуализация работы одной итерации алгоритма

Этот алгоритм был запущен 50 раз, и на каждой итерации выполнялись все необходимые условия, чтобы считать имитационную модель корректно описывающей заданную теоретическую модель CAOM.

Заданные начальные значения: для эффекта влияние количества общих связей – 1.5, для эффекта влияние количества исходящих связей – 1.8; интервалы времени – 2.5; интенсивности для акторов – 0.8; размерность сети – 50; коэффициент разреженности начального состояния сети – 0.75.

На рисунке 9 представлен график оценки параметров, где по оси абсцисс указан номер эксперимента, а по оси ординат — оценка параметра, полученная в результате эксперимента.

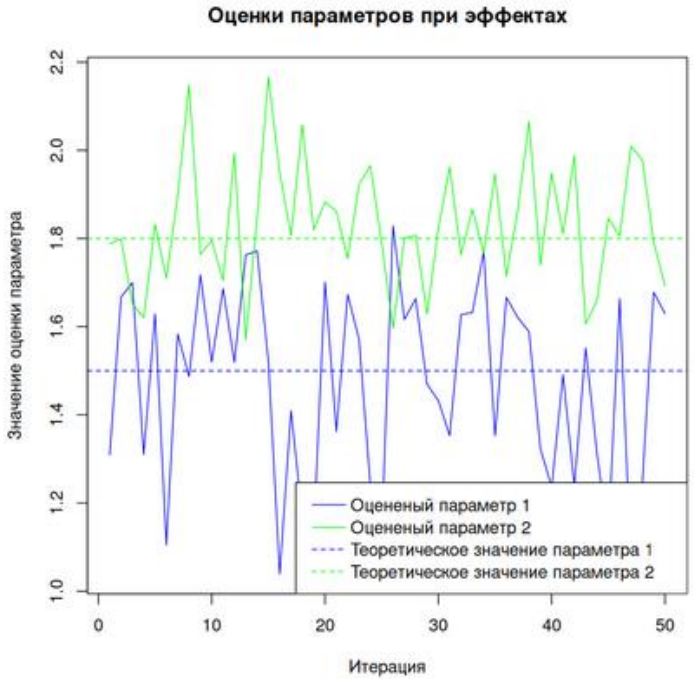

Рисунок 9 – График значений оценок параметров

На рисунке 10 представлены ядерные оценки плотности первого и второго параметра.

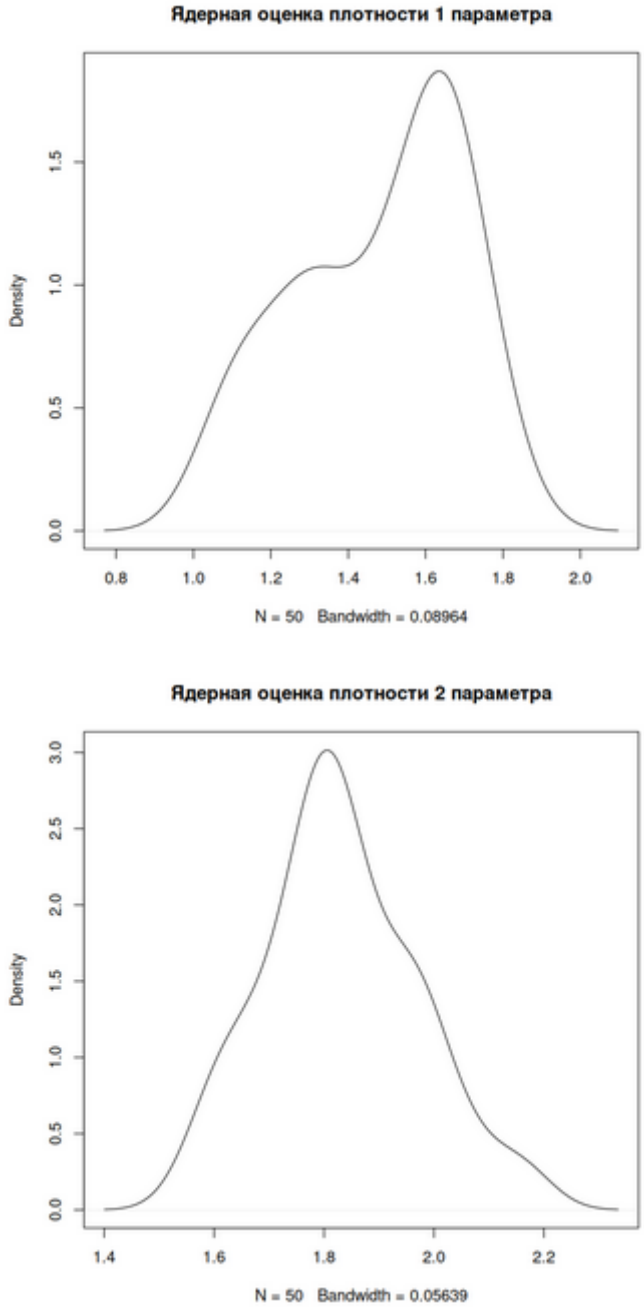

Рисунок 10 – Графики ядерных оценок плотности параметров

Выборочное математическое ожидание оценок параметра 1 эффекта *влияние количества общих связей* – 1.48, параметра 2 эффекта *влияние количества исходящих связей* – 1.83.

Средняя дисперсия оценок параметра 1 эффекта *влияние количества общих связей* – 0.267, параметра 2 эффекта *влияние количества исходящих связей* – 0.024.

Выборочная дисперсия оценок параметра 1 эффекта *влияние количества общих связей* – 0.047, параметра 2 эффекта *влияние количества исходящих связей* – 0.02.

ЗАКЛЮЧЕНИЕ

В рамках этой работы был изучен класс стохастических акторно ориентированных моделей. Поставленная цель, а именно построение и анализ стохастических акторно-ориентированных моделей на реальных данных и симуляция сетевой динамики была достигнута. Каждая из поставленных задач была выполнена.

Была исследована динамика сети научного сообщества ТГУ, и выявлены факторы влияющие на динамику сети. Так же была построена имитационная модель,

САОМ, является новым математическим методом, использующий методы и подходи из дисциплин ставших уже классическими. В САОМ используется теория потоков, но применение классических методов теории массового обслуживания или теории потоков затруднительно. Это связано с тем, что процесс, который моделируется, представляет собой сеть, и при переходе от одного состояния сети к другому одним из основных факторов является именно структура самой сети. Это требует использования инструментов теории графов. Также использование сетевых эффектов напоминает использование линейной регрессии, однако вместо предикторов используется поток принятий решений. Это усложняет анализ сетевых данных как задачи линейной регрессии. Сетевые структуры хранят в себе большое количество информации, а динамика этих сетевых структур позволяет узнать многое о виде сети, о акторах, и о том как внешние факторы влияют на рассматриваемую сеть.

Результаты, полученные в ходе анализа социальной сети научного соавторства Томского государственного университета, могут быть положены в основу дальнейшего исследования данной сети, а также создания новых социологических гипотез в отношении сети соавторства ТГУ. Данные исследования могут быть использованы для оптимизации научного сообщества. В наши дни все большую популярность набирают

междисциплинарные исследования, и в результате текущего исследования видно, что с увеличением стажа у авторов увеличивается вероятность междисциплинарной связи. Однако для дальнейшего исследования сети необходима разработка социологической теории о динамике сети и анализ поставленной теории. Полученные модели могут послужить не только для оптимизации взаимодействия научного сообщества, но и для оптимизации академического процесса студентов.

В дальнейшем планируется провести углубленное изучение сети научного сообщества ТГУ, включая проведение кластерного анализа и выделение существующих научных групп. Также предполагается рассмотреть динамику связных сетей, включая как динамику внутри каждой из научных групп, так и их влияние на остальные научные группы. Интерес представляет влияние научных конференций на степень кластеризации сети и их влияние на междисциплинарные исследования.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ И ЛИТЕРАТУРЫ

1. Докука С.В., Валеева Д.Р. Статистические модели для анализа динамики социальных сетей в исследованиях образования // Вопросы образования. - 2015. - №1. - С. 201-213.
2. Anuška Ferligoj, Luka Kronegger, Franc Mali, Tom A B Snijders, Patrick Doreian Scientific collaboration dynamics in a national scientific system // Scientometrics. - 2015. - №104. - С. 985–1012.
3. Snijders, T.A.B., & Pickup, M. (2016). Stochastic Actor-Oriented Models for Network Dynamics. Retrieved June 10, 2016.
4. Manual for RSiena / M. R. Ruth. — Текст : электронный // University of Oxford: Department of Statistics; Nuffield College University of Groningen: Department of Sociology : [сайт]. — URL: https://www.stats.ox.ac.uk/~snijders/siena/RSiena_Manual.pdf (дата обращения: 14.01.2024).
5. Snijders, T. A. B. The Statistical Evaluation of Social Network Dynamics // Sociological Methodology. - 2001. - №31. - С. 361-395.
6. Siena algorithms // Department of Statistics – University of Oxford URL: https://www.stats.ox.ac.uk/~snijders/siena/Siena_algorithms.pdf (дата обращения: 06.01.2024).
7. Tom A.B. Snijders, Gerhard G. van de Bunt, Christian E.G. Steglich Introduction to Stochastic Actor-Based Models for Network Dynamics // Social Networks. 2010. №32. С. 44-60.
8. Tom Broekel, Pierre-Alexandre Balland, Martijn Burger, Frank Oort Modeling knowledge networks in economic geography: a discussion of four methods // The Annals of Regional Science. - 2014. - №53. - С. 423-452.