DOI: 10.25728/mlsd.2023.1598

МОДЕЛИРОВАНИЕ ПРОЦЕССОВ ИНФОРМАЦИОННОГО ПРОТИВОБОРСТВА: ТЕОРИЯ АРГУМЕНТНЫХ ВЗАИМОДЕЙСТВИЙ И ФЕЙКОВЫЕ НОВОСТИ

Антонов А.В., Козицин И.В.
Институт проблем управления им. В.А. Трапезникова РАН, Москва, Россия
antonov.aleksej@phystech.edu, kozitsin.ivan@mail.ru

Аннотация. В данной работе строится оригинальная агентная модель, позволяющая исследовать процессы, связанные с распространением фейковых новостей в онлайн среде и их влиянием на общественное мнение в условиях, когда два информационных источника являются противоборствующими сторонами информационного конфликта, борющимися за мнения обычных пользователей.

Ключевые слова: теория аргументных взаимодействий, фейковые новости, модели динамики мнений, алгоритмы ранжирования.

Введение

Интенсификация информационный процессов, вызванная стремительным развитием Интернета и онлайновых сетей способствует продвижению фейковых (ложных) новостей онлайн среде, несущих значительную угрозу для устойчивого развития общества [1]. Использование фейкового контента позволяет некоторым заинтересованным лицам оказывать значительное влияние на общественное мнение, поскольку, будучи не скованными необходимостью опираться на достоверные факты и события, фейковые новости зачастую более виральны и обладают бОльшей убеждающей силой [2].

Агентные модели социального влияния – мощный инструмент из арсенала математического моделирования, позволяющий изучать каким образом формируются взгляды людей под действием социального влияния и той информации, которую они получают извне, в том числе из сети Интернет [3, 4]. Однако в такого рода моделях мнения агентов как правило формализуются достаточно упрощенным образом – при помощи скаляров или векторов, описывающих позицию человека относительно одного или нескольких актуальных вопросов повестки дня [5]. Такой подход не позволяет учитывать правдивость информации, которой индивиды обмениваются друг с другом.

Продуктивным видится использование более “атомарного” подхода, при котором рассматривается двухуровневое когнитивное устройство человека. На нижнем уровне находятся аргументы-факты, в которые верит агент и на основании которых формируется его позиция, которая располагается на втором, более высоком уровне. Такая конструкция была предложена в Теории аргументных взаимодействий (argument-based communication theory) [6–10] и получила широкую известность, поскольку при помощи нее удалось описать явление поляризации общественного мнения без использования механизма репульсивного влияния [11, 12].

В данной работе строится оригинальная агентная модель, позволяющая исследовать процессы, связанные с распространением фейковых новостей в онлайн среде в условиях, когда два информационных источника являются противоборствующими сторонами информационного конфликта и продвигают противоположные точки зрения относительно острого социального-экономического вопроса [13–15]. Модель позволяет изучать влияние активности агентов, устройства алгоритмов ранжирования и топологии социальной сети на динамику системы (в том числе, на динамику мнений агентов). В этом смысле данная модель является продолжением модели, предложенной в работе [16]. Кроме того, по сравнению с работой [16], в модель добавлена динамика социального графа, обусловленная желанием агентов быть подписанными на тот информационный источник, повестка которого когерентна их взглядам [17]. В отличие от [16], в данной статье мнения агентов рассматриваются через призму Теории аргументных взаимодействий. При этом, в данной статье учитывается достоверность аргументов, на основании которых строятся мнения агентов, а также их сила (определяемая как вклад аргумента в итоговое мнение агента). Помимо этого, предлагается рассматривать динамически пополняемое множество аргументов. Такая постановка является более точным описанием реального мира, в котором новостные источники ежедневно транслируют истории и факты, выступающие базисом для формирования новых аргументов.

Для построенной модели реализуется программная оболочка на языке Python, модель исследуется при помощи численных экспериментов, рассматриваются два базовых сценария информационного противоборства.

Данная работа является расширенной версией публикации, представленной на Школе конференции УБС-2022. По сравнению с конференционной публикацией: исправлено описание модели, написан и

выложен в открытый доступ программный код (см. далее), проведены численные эксперименты, представлены и проинтерпретированы их результаты.

Дальнейшее изложение структурировано следующим образом. В разделе 2 представлена основополагающая модель информационных взаимодействий. Раздел 3 содержит описание дизайна имитационных экспериментов. В разделе 4 представлены результаты имитационных экспериментов. Раздел 5 содержит вывод; также в это разделе приведены направления для дальнейших исследований.

1. Модель

1.1. Краткий обзор

Модель призвана имитировать информационные процессы внутри некоторой онлайновой сети, в результате которых меняются мнения пользователей. Рассматривается закрытая система из $N + 2$ агентов, связанных социальной сетью $G = (V, E)$. Множество V описывает вершины сети, а E – связи между ними. Вершины индексируются от 1 до $N + 2$. Вершины с индексами $N + 1$ и $N + 2$ – это агенты, соответствующие акторам – сторонам ин-формационного конфликта. Содержательно это могут быть информационные источники в реальных онлайновых сетях (далее этих агентов будем называть информационными источниками), а остальные агенты соответствуют обычным пользователям (далее – просто агенты).

Связи между агентами являются двусторонними, а ребра, соединяющие агентов и информационные источники – односторонние, направленные в сторону от агентов. Информационные источники не могут быть связаны друг с другом. Однако возможны ситуации, когда один агент подписан одновременно на оба информационных источника. Связи между агентами заморожены, однако связи между агентами и информационными источниками могу изменяться со временем.

Время в модели дискретно: $t = 1, 2, 3, \ldots, T$. Каждый ход случайным образом выбирается один из информационных источников, который может опубликовать пост с некоторой вероятностью. Если выбранный информационный источник не публикует пост, то тогда действует выбранный случайным образом агент, который может осуществить одно из некоторого списка наперед заданных действий. Также каждые N тактов времени (один шаг Монте-Карло [18]) активируется алгоритм проверки достоверности новостей. Таким образом в рамках одного хода может произойти одновременно два события: ход информационного источника или агента, а также проверка достоверности.

1.2. Аргументы и мнения

Ключевым элементом модели являются аргументы – кванты информации, которые участвуют в формировании мнений на основании так называемой оценочной структуры (evaluative structure) [7]. Для обозначения аргументов используется символ “a”. Каждый аргумент a характеризуется триплетом $a = (s, w, f)$, где s определяет валентность аргумента (+1 или -1), w задает вес данного аргумента в оценочной структуре (принимает значения от нуля включительно до плюс бесконечности), а f является индикатором того, что данный аргумент является фейком (к примеру, основан на ложном факте) и принимает значения +1 (фейк) или 0 (не фейк). Каждый агент характеризуется кортежем дуплетов аргумент-индикатор: $A = ((a_1, e_1), \ldots, (a_m, e_m))$. В дуплете (a_i, e_i) индикатор e_i описывают, принят ли аргумент a_i ($e_i = 1$) агентом или нет ($e_i = 0$). Мнение агента строится на основании принятых аргументов кортежа:

$$o = \sum_k e_k s_k w_k. \tag{1}$$

Конструкция (1) задает отображение из множества аргументов на вещественную ось и является ключевой в Теории аргументных взаимодействий. Один из вкладов данной статьи состоит во введении весов аргументов, отвечающих за их вклад в итоговое мнение, а также их достоверности. У разных агентов кортежи могут иметь различную длину. В начальный момент времени все кортежи агентов пусты и мнения агентов тем самым равны нулю. Ниже будет приведен механизм пополнения кортежа и определения значений индикаторов.

1.3. Информационные источники

Ход информационного источника состоит в публикации поста. Каждый пост p в момент времени t характеризуется квартетом $p = (i, t_0, a, l)$, где i – индекс автора поста, t_0 – время публикации (такт), a – аргумент, который присутствует в посте и l – число лайков, которые собрал пост к моменту времени t. В момент создания пост не имеет лайков: $l(t_0) = 0$. Аргумент поста генерируется при создании

последнего информационным источником как случайная величина. Информационные источники генерируют новые (уникальные) аргументы. При создании поста информационный источник $i \in \{N + 1, N + 2\}$ обращается к одномерному распределению F_i. На основании этого распределения генерируется случайная величина ξ_i. Эта случайная величина порождает аргумент, инициализируемый как $a = (\mathrm{sgn}(\xi_i), |\xi_i|, f)$. Если не указано обратное, то считаем $f = 0$. Информационный источник i публикует пост в свой ход с вероятностью α_i.

Цели, которые преследуют информационные источники, состоят в (i) изменении мнений агентов и (по возможности) (ii) максимизации числа подписчиков. Далее для ясности будем считать, что информационный источник с индексом $N + 1$ тяготеет к отрицательной части пространства мнений, а информационный источник с индексом $N + 2$ – к положительной. Математически это выражается в том, что распределение F_{N+2} находится правее F_{N+1}.

Информационный источник $i \in \{N + 1, N + 2\}$ публикует пост с вероятностью α_i в свой ход.

1.4. Агенты и их действия

Каждый агент при активации может выполнить одно из следующих трех действий: (i) Опубликовать пост; (ii) Ознакомиться с новостной лентой; (iii) Блуждать по онлайновой сети (англ. выражение surf the Internet). Вероятности каждого действия α, β, и γ соответственно, где $\alpha + \beta + \gamma = 1$. Опишем, что происходит в рамках каждого из этих действий.

Действие “опубликовать пост”. Данное действие идентично тому, что делают информационные источники с одной поправкой: агенты, в отличие от информационных источников, не генерируют новые аргументы, а оперируют теми, которые входят в их картежи. Только информационные источники могут создавать новые аргументы. Более точно, агент, кортеж аргументов которого описывается $A = ((a_1, e_1), \ldots, (a_m, e_m))$, для создания поста использует последний принятый элемент кортежа a_j (с $e_j = 1$). Если кортеж аргументов агента пуст или в нем нет аргументов с флагом $e = 1$, то тогда агент не может выполнить данное действие, и ничего не происходит. Последняя ситуация является достаточно “экзотичной”, интерес представляет сама возможность существования такой конфигурации кортежа аргументом.

Действие “ознакомиться с новостной лентой”. В рамках данного действия агент просматривает посты, опубликованные его соседями по социальной сети (под соседями подразумеваются все инцидентные вершины независимо от типа связи (двусторонняя / односторонняя)). Агент имеет доступ к постам, опубликованным с момента его последней активации: если агент последний раз активировался в такт времени t, то тогда, будучи активированным в такт T, ему будут видны посты, опубликованные инцидентными вершинами в моменты времени $t + 1, t + 2, \ldots, T - 1$ (если таковые найдутся). Множество таких постов обозначим $P_{t,T}$. Из этого множества Алгоритм ранжирования R выбирает один пост $p = R(P_{t,T})$, который затем просматривается агентом. Если множество $P_{t,T}$ пустое, то ничего не происходит.

В рамках модели предлагается рассмотреть два алгоритма ранжирования. Первый (R_1) из множества постов $P_{t,T}$ выбирает самый новый. Второй алгоритм (R_2) предлагает для агента пост с наибольшим числом лайков. Если таковых несколько, то выбирается самый новый пост.

Действие “блуждание по сети”. При совершении данного действия агент посещает аккаунт случайно выбранного информационного источника (по умолчанию, считаем, что выбор информационных источников равновероятен, однако в дальнейшим целесообразно рассмотреть более точные приближения реальных процессов, в рамках которых вероятность посетить аккаунт положительно связана с числом подписчиков данного аккаунта). Данное действие не зависит от того, подписан ли агент на конкретный информационный источник или нет. При посещении аккаунта агент видит последний пост, опубликованный данным информационным источником.

1.5. Просмотр поста агентом

Данное событие является основным механизмом передачи информации и, тем самым, оказания влияния в рамках модели. Просмотр поста агентом может возникнуть в рамках действий “Ознакомиться с новостной лентой” и “Блуждание по сети”. Пусть агент, характеризуемый кортежем аргументов $A = ((a_1, e_1), \ldots, (a_m, e_m))$ и мнением $o = \sum_k e_k s_k w_k$ наблюдает пост p, несущий аргумент a (с валентностью s и весом w).

Возможны три ситуации.

Ситуация 1. Предположим, что аргумент a, содержащийся в посте p не входит в кортеж A. В таком случае агент может либо принять аргумент, добавив в кортеж с флагом $e = 1$, либо отвергнуть – в

последнем случае кортеж останется неизменным. В рамках первого случая кортеж примет вид $A = ((a_1, e_1), \ldots, (a_m, e_m), (a, 1))$, а мнение агента станет равно $o = \sum_k e_k s_k w_k + s w$.

Ситуация 2. Теперь рассмотрим ситуацию, когда аргумент a входит в кортеж A с флагом $e = 1$. В этом случае возможны два варианта событий: аргумент останется с флагом $e = 1$ (аргумент остается принятым), или же будет вычеркнут из кортежа: $e = 0$ (аргумент отвергнут).

Ситуация 3. В случае если аргумент входит в кортеж с флагом $e = 0$, то также возможны два варианта событий: аргумент останется с флагом $e = 0$ (аргумент остается отвергнутым), или же будет принят агентом: $e = 1$ (аргумент принят).

Таким образом, аргумент может попасть в кортеж только с флагом $e = 1$. Значение флага может поменяться только при переоценке аргумента. Если в результате взаимодействия агента с постом аргумент принят (Ситуации 1, 3) или остается принятым (Ситуация 2), то тогда агент может с вероятностью ζ поставить лайк данному посту. Если, кроме того, пост опубликован информационным источником, на который агент не подписан, то агент может на него подписаться с вероятностью η. Если же в результате взаимодействия агента с постом аргумент отвергнут (Ситуации 1, 2) или остается отвергнутым (Ситуация 3), то если пост был опубликован информационным источником, на который подписан агент, то тогда агент может отписаться от данного информационного источника с вероятностью θ.

1.6. Принятие / непринятие аргумента агентом

Чтобы задать правило, на основании которого агенты принимают или отвергают аргументы, введем величину, описывающую когерентность когнитивной конфигурации кортеж – мнение:

$$C[A] = \sum_k E_k s_k w_k \sum_k e_k s_k w_k, \tag{2}$$

где $E_k = 2e_k - 1$ – величина, принимающая значения +1 или -1. Когерентность (2) растет в случае, если кортеж пополняется аргументами, коррелирующими с текущим мнением агента, а значит коррелирующими с аргументами, которые были приняты агентом ранее. Когерентность также положительно связана с отвергнутыми аргументами, отрицательно коррелирующими с текущим мнением.

Эмпирические исследования свидетельствуют о том, что люди стремятся получать информацию извне, которая подтверждает их взгляды (смещенная обработка информации – biased processing). В связи с этим, следуя работе [3], будем считать, что, обрабатывая новый для себя аргумент a, агент ориентируются на то, как изменится когерентность когнитивной конфигурации в случае его принятия:

$$V = C[\{(a_1, e_1), \ldots, (a_m, e_m), (a, 1)\}] - C[\{(a_1, e_1), \ldots, (a_m, e_m)\}] \tag{3}$$

Вероятность принятия аргумента a агентом с кортежем A описывается величиной λ, которая определяется через логистическую модель:

$$\lambda = \frac{1}{1 + e^{-\mu V}}. \tag{4}$$

В формуле (4) величина $\mu \geq 0$ регулирует силу эффекта смещенной обработки информации (при $\mu = 0$ вероятность принятия/непринятия аргумента не зависит от того, как при этом изменится когнитивная конфигурация), а V определяется выражением (3).

1.7. Фейковые аргументы и механизм проверки достоверности информации

Информационные источники, которые генерируют новые аргументы и имплементируют их в свои посты, могут повлиять на свойства аргументов, поступившись их “правдивостью”. Более точно, информационный источник i вместо распределения F_i может использовать другое вероятностное распределение G_i. В результате сформированный аргумент будет иметь следующую структуру: $a = (\mathrm{sgn}(\xi_i), |\xi_i|, 1)$, то есть будет фейковым (основанным на фейковой информации). Предполагается, что распределение G_i лучше отвечает целям информационного источника, чем распределение F_i (G_{N+1} расположено левее F_{N+1} и G_{N+2} расположено правее F_{N+2}). Сформированные таки образом фейковые аргументы могут быть приняты агентами, которые далее могут донести их до своих знакомых.

Если информационный источник i публикует пост в свой ход, то он будет фейковым с вероятностью A_i. Таким образом величины A_{N+1} и A_{N+2} регулируют число фейковых постов (и аргументов) в онлайновой сети.

В моделируемой онлайновой сети работает механизм разоблачения фейков. С некоторой периодичностью (по умолчанию – раз в N ходов) происходит процедура проверки достоверности аргументов. В рамках данной процедуре проверке подвергаются посты, опубликованные информационными источниками за последние (по умолчанию) N тактов. Каждый такой пост p с

вероятностью v проходит проверку путем установления достоверности скрытого в нем аргумента a. Если аргумент a достоверен ($f = 0$), то тогда ничего не происходит. В противном случае (аргумент основан на ложной информации – $f = 1$) онлайновая сеть официально признает аргумент фейковым. С точки зрения модели происходит процедура обновления кортежей агентов: для каждого агента напротив данного аргумента в его кортеже (если он присутствует в кортеже) устанавливается флаг $e = 0$. Помимо этого, если агент подписан на информационный источник, который является автором данного фейкового поста, то тогда с вероятностью φ агент отписывается от него.

1.8. Обзор параметром модели и программная реализация

Модель включает в себя достаточно большое число гиперпараметров – переменных, которые являются внешними по отношению к модели и которые необходимо задать заранее. В таблице 1 приведена их краткая сводка.

Модель реализована на языке Python для проведения численных экспериментов. Актуальная версия программного кода доступна по адресу https://doi.org/10.7910/DVN/WNUFFE. Программный код снабжен процедурой для сохранения результатов эксперимента, что дает возможность накапливать экспериментальную базу, повторяя один и тот же эксперимент несколько раз для одних и тех же значений гиперпараметров – иными словами, это дает возможность исследовать модель методом Монте-Карло.

Таблица 1. Параметры модели и их описание

Параметр	**Описание**
T	Число итераций
N	Число агентов в системе (+ 2) информационных источника
$G = (V, E)$	Социальный граф, соединяющий агентов и информационные источники.
Начальное распределение мнений (структура кортежей агентов при $t = 1$)	
α_{N+1}, α_{N+2}	Вероятности публикации поста информационными источниками.
A_{N+1}, A_{N+2}	Вероятности того, что опубликованные информационными источниками посты будут включать фейковый аргумент.
α, β, γ	Вероятности действий “Публикация поста”, “Ознакомиться с новостной лентой” и “Блуждать по сети”, совершаемых агентом в свой ход. Данные действия образуют полное пространство событий, в связи с чем $\alpha + \beta + \gamma = 1$.
F_{N+1}, F_{N+2}	Вероятностные распределения, на основании которых вероятностные источники генерируют достоверные аргументы (аргументы, основанные на достоверной информации).
G_{N+1}, G_{N+2}	Вероятностные распределения, на основании которых вероятностные источники генерируют фейковые аргументы (аргументы, основанные на ложной информации).
R_1, R_2	Алгоритмы ранжирования. Алгоритм R_1 выбирает для пользователя тот пост из новостной ленты, который был позднее всего опубликован, а R_2 отбирает пост с наибольшим числом лайков.
ζ	Вероятность поставить лайк посту, аргумент которого был принят агентом.
η	Вероятность подписаться на информационный источник, опубликовавший пост, аргумент которого был принят агентом.
θ	Вероятность отписаться от информационного источника, опубликовавшего пост, аргумент которого был отвергнут агентом.
μ	Сила эффекта смещенной обработки информации
v	Вероятность проверки поста в рамках процедуры проверки достоверности информации
φ	Вероятность отписаться от информационного источника, пост которого был признан фейковым

2. Численные эксперименты

2.1. Общий подход

В рамках модели происходит (i) динамика мнений агентов и (ii) динамика социального графа. Коэволюция этих двух процессов порождает достаточно сложную феноменологию, описание которой

в рамках аналитического подхода затруднено. В связи с этим целесообразно исследовать модель с помощью численных экспериментов, варьируя параметры, представленные в таблице 1. Число таких параметров слишком велико для проведения исчерпывающего анализа (покрывающего все возможные комбинации параметров). Вместе с тем некоторые параметры могут быть идентифицированы на основании нашего опыта, представлений экспертов или эмпирических данных.

В данной работе мы остановимся на двух базовых сценариях информационного противоборства, зафиксировав большинство гиперпараметров в соответствии с нашими базовыми представлениями об устройстве объекта исследования. Отметим, что цель экспериментов – продемонстрировать работу модели, проанализировав некоторые содержательные ситуации, которые могут возникнуть в реальной жизни. Полномасштабный анализ феноменологии модели будет проведен в дальнейших исследованиях. В таблице 2 приведены значения гиперпараметров, которые НЕ будут варьироваться в рамках имитационных экспериментов (за исключением тех случаев, когда это оговорено отдельно). Было принято решение использовать граф Ваттса-Строгатца в качестве модели социальной сети, поскольку в данном графе распределение степеней вершин гомогенно. Таким образом информационные источники в начальный момент находятся примерно в одинаковых условиях. Если бы распределение степеней вершин было гетерогенно (как, например, в модели Барабаши-Альберт [19]), то тогда необходимо было бы контролировать степени вершин информационных источников в начальный момент времени.

Таблица 2. Значения "замороженных" гиперпараметров

Параметр	**Значение**
T	150,000 (пилотные эксперименты выявили, что данного числа итераций достаточно для понимания поведения динамики модели)
N	200
$G = (V, E)$	Граф Ваттса-Строгатца [20] на $N + 2$ вершин с параметрами $k = 7$ и $p = 0.2$. Такой граф формируется на основании кольца из вершин, каждая из которых (в том числе) соединена с k ближайшими. Далее каждое ребро с вероятностью p случайным образом меняет инцидентные вершины. Образованный таким образом граф обладает высоким значением коэффициента транзитивности (~ 0.172) и характеризуется небольшой средней длиной пути (~ 3, что почти в два раза меньше, чем логарифм от числа вершин – такой граф называется "мир тесен")
Начальное распределение мнений (структура кортежей агентов при $t = 1$)	В начальный момент времени кортежи агентов пусты. Таким образом все мнения равны нулю
α_{N+1}, α_{N+2}	$\alpha_{N+1} = \alpha_{N+2} = 0.1$
A_{N+1}	0 (информационный источник с индексом $N + 1$ всегда публикует достоверные посты)
α, β, γ	0.2, 0.6, 0.2 – агенты отдают предпочтение чтению новостной ленты
F_{N+1}, F_{N+2}	Нормальные распределения со средними -2 и 2 и стандартным отклонением 1
G_{N+1}, G_{N+2}	Нормальные распределения со средними -7 и 7 и стандартным отклонением 3 (см. рисунок 1)
ζ	0.5
η	0.1
θ	0.1
μ	1
φ	0.2

Те примеры, которые будут рассмотрены, далее обыгрывают ситуацию, когда один информационный источник (без ограничений общности - с индексом $N + 1$) публикует достоверные посты, а второй поступается честностью той информации, которую он публикует для достижения своих целей. Среди прочего могут возникнуть два вопроса: как должен вести себя второй информационный источник и каким образом ему может противодействовать социальная сеть в рамках институциональных мер.

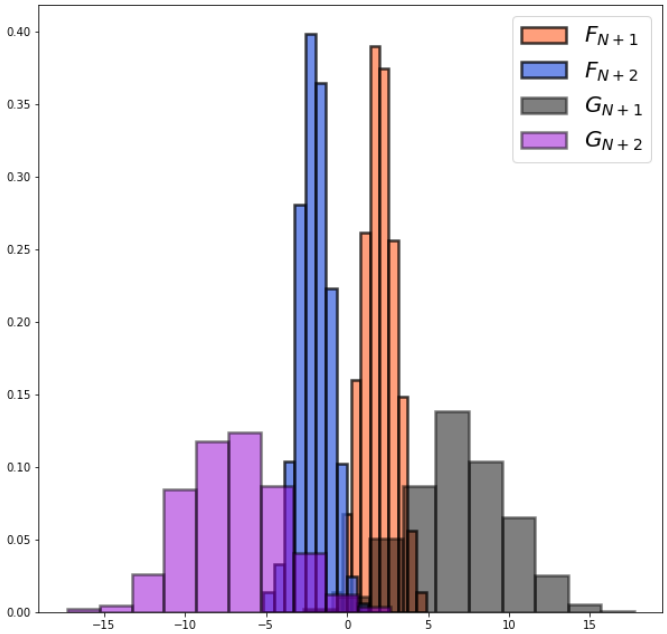

Рис. 1. Распределения, используемые для генерации аргументов в рамках имитационных экспериментов

С целью построения статистически значимых заключений, для каждого набора гиперпараметров было проведено 40 независимых испытаний.

2.2. Сценарий 1

В рамках данного сценария будет зафиксирована вероятность проверки подлинности постов $v = 0.1$. Варьироваться будет величина A_{N+2}, принимая значения из множества {0, 0.01, 0.05, 0.1, 0.5}. Таким образом будет изучено влияние вероятности публикации фейковых постов вторым информационным источником на исход информационного конфликта – какую выгоду может извлечь для себя второй информационный источник, публикуя заведомо недостоверную информацию при заданном устройстве онлайновой сети?

2.3. Сценарий 2

Данный сценарий фиксирует вероятность публикации фейковых постов вторым информационным источником $A_{N+2} = 0.5$ и рассматривает различные значения v из множества {0, 0.01, 0.05, 0.1, 0.5}. Иными словами, мы изучаем, каким образом широта охвата постов механизмом проверки подлинности влияет на исход информационного конфликта при условии, что стратегия второго информационного источника (под стратегией мы понимаем вероятность публикации фейковых постов) фиксирована.

3. Результаты

Все результаты, представленные в данном разделе получены при алгоритме ранжирования R_1. При использовании алгоритма R_2 результаты остаются теми же как на качественном, так и на количественном уровнях.

3.1. Сценарий 1

Анализ результатов экспериментов выявил, что с ростом вероятности публикации фейковых постов, мнения агентов все сильнее сдвигаются в положительную область пространства мнений, соответствующую, напомним, интересам второго информационного источника (см. рис. 2). При этом, что важно, растет как число агентов с положительным мнением, так и магнитуда их мнений. Чтобы добиться 65 процентов голосов, второй информационный источник должен лгать в пятидесяти процентах случаев.

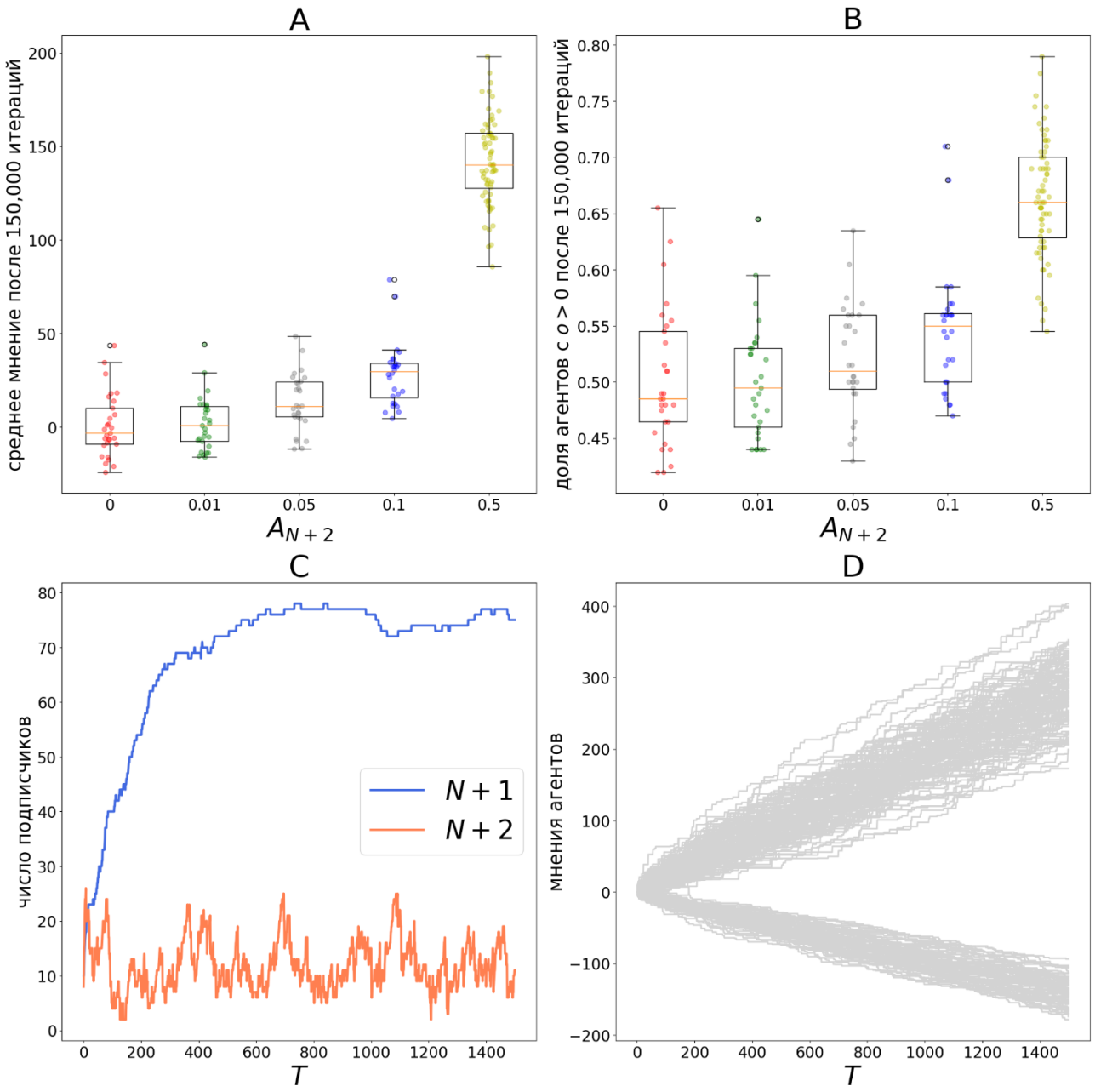

Рис. 2. (A) Среднее мнение агентов после 150,000 итераций как функция A_{N+2}. (B) Доля агентов с положительным мнением после 150,000 итераций как функция A_{N+2}. (C) Типичный график численностей подписчиков первого и второго информационных источников для A_{N+2} = 0.5. (D) Динамика мнений агентов для того же эксперимента, что и в пункте (C)

Охарактеризуем типичный вид динамики системы в рамках данного сценария. Из рис. 1 видно, что агенты формируют два кластера, поляризующиеся вокруг позиций, продвигаемых информационными источниками. При этом аудитория второго информационного источника колеблется, так как каждые 200 итераций (после проверки подлинности информации, содержащейся в постах) часть агентов отписываются от него. Что интересно, с ростом вероятности публикации фейковых постов, аудитория второго информационного источника уменьшается (см. рис. 3), но это не мешает ему достигать своей основной цели, связанной с воздействием на общественное мнение. Более того, первая (влияние на общественное мнение) и вторая (максимизация числа подписчиков) цели второго информационного источника в некотором роде являются взаимоисключающими - чем меньше значение A_{N+2}, тем больше подписчиков имеет второй информационный источник, но тем менее выражен сдвиг общественного мнения в правую сторону спектра мнений. Объяснение данному эффекту следующее – с ростом вероятности публикации фейковых новостей вторым информационным источником, его аудитория уменьшается, но становится более радикальной, что позволяет информационному источнику опосредованно влиять на друзей своих подписчиков.

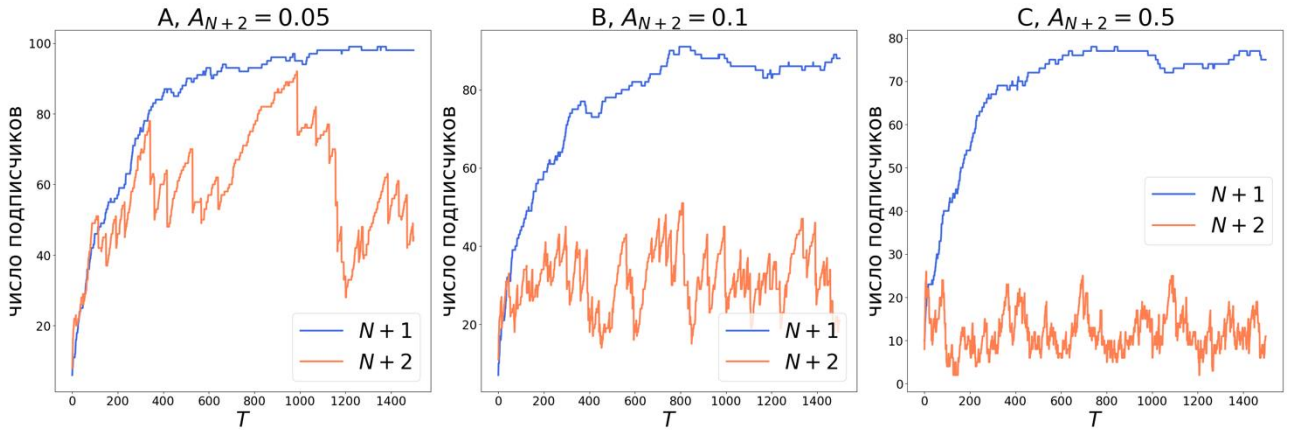

Рис. 3. Динамика численностей подписчиков информационных источников для разных значений A_{N+2}

Важно отметить, что такого рода динамика характерна именно для значения параметра смещенной обработки информации $\mu = 1$. К примеру, при $\mu = 0$ (эффект смещенной обработки информации отсутствует) поляризация не наблюдается и распределение мнений агентов близко к нормальному – см. рис. 4. Иными словами, наличие эффекта смещенной обработки информации является необходимым условием для формирования эхо-камер и поляризации мнений (в рамках рассматриваемых сценариев).

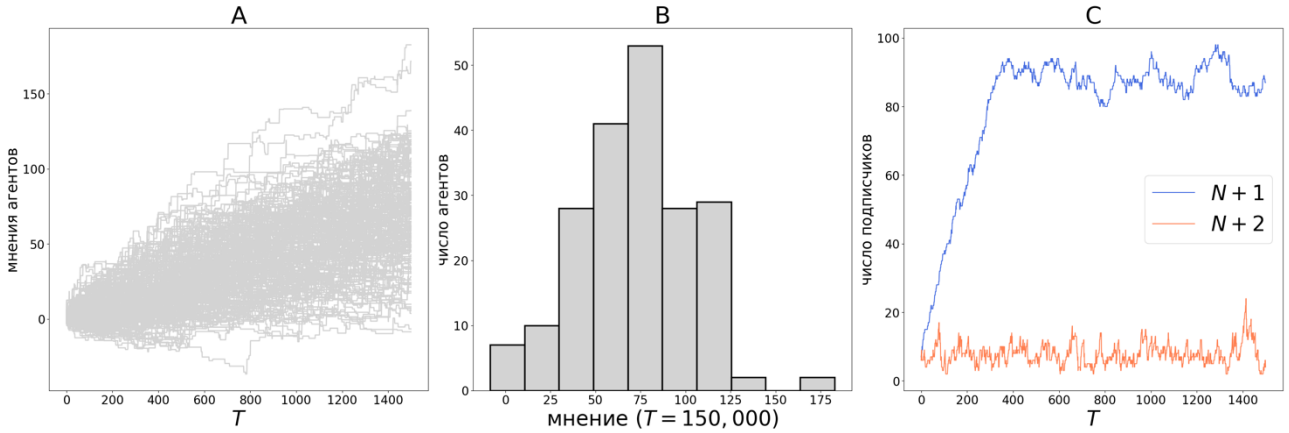

Рис. 4. Поведение модели при $\mu = 0$ и $A_{N+2} = 0.5$. (A) Типичный вид динамики мнений агентов. (B) Типичный вид распределения мнений агентов после 150,000 итераций. (C) Типичный график численностей подписчиков первого и второго информационных источников. В пунктах (B) и (C) рассматривается тот же эксперимент, что и в пункте (A))

3.2. Сценарий 2

Рис. 5 демонстрирует, что с ростом широты охвата постов механизмом проверки подлинности общественное мнение сдвигается в левую сторону пространства мнений, нивелируя то преимущество, которое было у второго информационного источника за счет использования фейков ($A_{N+2} = 0.5$). Чтобы сместить среднее мнение к нулю, нужно проверять на достоверность порядка половины постов $v \sim 0.5$ (что чрезвычайно много для реальной онлайновой сети). Однако, чтобы добиться равенства числа агентов с положительными и отрицательными мнениями, достаточно проверки порядка процентов постов ($v \sim 0.4$). Проверка одного процента постов ($v \sim 0.01$) гарантирует, что второй информационный источник будет иметь преимущество в терминах числа подписчиков. Однако небольшое увеличение охвата проверки ($v \sim 0.05$) нивелирует это преимущество.

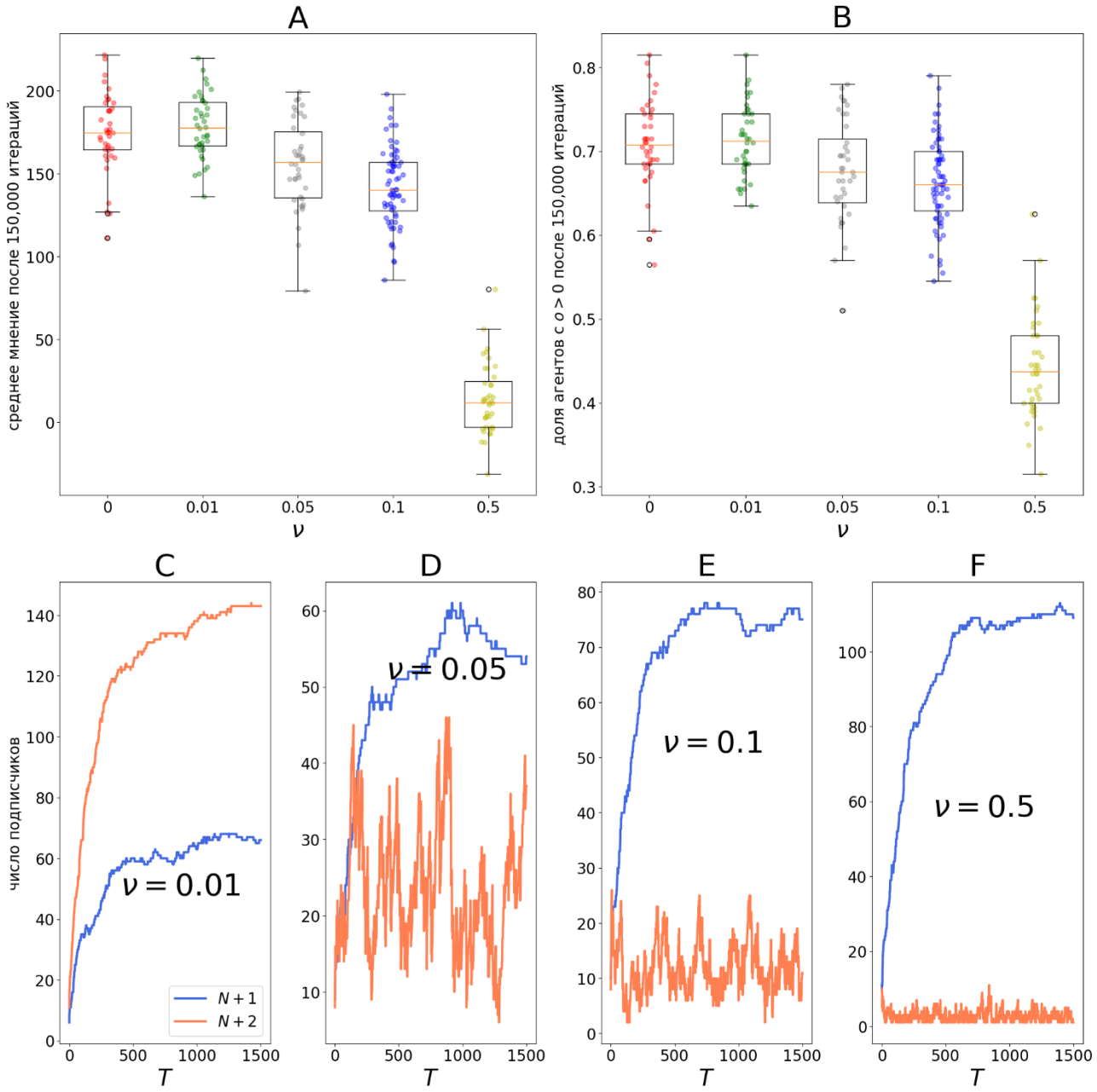

Рис. 5. Поведение модели при $A_{N+2} = 0.5$. (A) Среднее мнение агентов после 150,000 итераций как функция ν. (B) Доля агентов с положительным мнением после 150,000 итераций как функция ν. (C)–(F) Типичные графики зависимости численностей подписчиков первого и второго информационных источников от значений ν

4. Заключение

В данной статье была предложена оригинальная агентная модель, позволяющая исследовать процессы, связанные с распространением фейковых новостей в онлайн среде в условиях, когда два информационных источника являются противоборствующими сторонами информационного конфликта и продвигают противоположные точки зрения относительно острого социально-экономического вопроса. Модель позволяет изучать влияние активности агентов, устройства алгоритмов ранжирования и топологии социальной сети на динамику системы (в том числе, на динамику мнений самих агентов).

Мнения агентов в модели строятся на основании аргументов, которыми они оперируют (подход взят из Теории аргументных взаимодействий). Аргументы могут быть достоверными, а могут быть основаны на ложных фактах (фейковые аргументы). В систему искусственно введены два специальных агента – информационные источники, которые являются противоборствующими сторонами информационного конфликта. Также в модель имплементирована динамика социального графа,

обусловленная желанием агентов быть подписанными на тот информационный источник, повестка которого когерентна их взглядам. Информационные источники публикуют посты, в которых возникают новые аргументы (новые факты -> новые аргументы), перенимаемые затем агентами. Один из информационных источников специально поступается правдивостью постов с целью сделать увеличить их убеждающую силу. Однако онлайновая сеть борется с фейковыми постами путем точечной проверки. При обнаружении таковых их автор несет наказание – его подписчики отписываются от него, изменяя при этом свои взгляды (в связи с утратой доверия фейковым аргументам).

Для построенной модели реализована программная оболочка на языке Python, модель была исследована при помощи численных экспериментов, рассматривались два базовых сценария информационного противоборства. В рамках первого сценария изучалось влияние вероятности публикации фейковых постов вторым информационным источником на исход информационного конфликта – какую выгоду может извлечь для себя второй информационный источник, публикуя заведомо недостоверную информацию при заданном устройстве онлайновой сети. Второй сценарий был посвящен двойственному вопросу. А именно, каким образом широта охвата постов механизмом проверки подлинности влияет на исход информационного конфликта при условии, что стратегия второго информационного источника (под стратегией мы понимаем вероятность публикации фейковых постов) фиксирована.

Результаты имитационных экспериментов выявили, что при условии наличия эффекта смещенной обработки информации мнения агентов поляризуются вокруг позиций, продвигаемых информационными источниками. При этом второй информационный источник, публикуя фейковые посты, теряет аудиторию, однако, несмотря на это, значительно влияет на общественное мнение за счет небольшого ядра своих радикально настроенных подписчиков, которые транслируют позицию данного информационного источника своим друзьям.

Дальнейшие исследования могут быть связаны с анализом влияния топологии сети (в данной статье рассматривалась только одна топология – модель Ваттса-Строгатца) и различных типов алгоритмов ранжирования (в данной статье рассматривался тривиальный алгоритм, который среди множества постов новостной ленты выдает тот, который был позже всего опубликован).

Литература

1. *Chen W., Pachec, D., Yang K.C., Menczer, F.* (2021). Neutral bots probe political bias on social media // Nature communications. – 2021. – Vol. 12. – №. 1. – P. 1-10.
2. *Del Vicario M., Bessi A., Zollo F., Petroni F., Scala A., Caldarelli G., Quattrociocchi W.* (2016). The spreading of misinformation online // Proceedings of the National Academy of Sciences. – 2016. – Vol. 113. – №. 3. – P. 554-559.
3. *Chkartishvili A.G.* Social Networks: Models of information influence, control, and confrontation / A.G. Chkartishvili, D.A. Gubanov, D.A. Novikov – Springer, 2018. – Vol. 189. – 157 p.
4. *Flache A., Mäs M., Feliciani T., Chattoe-Brown E., Deffuant G., Huet S., Lorenz J.* Models of social influence: Towards the next frontiers // Journal of Artificial Societies and Social Simulation. – 2017. – Vol. 20. – №. 4. 10.18564/jasss.3521
5. *DeGroot M.H.* Reaching a consensus // Journal of the American Statistical association. – 1974. – Vol. 69. – №. 345. – P. 118-121.
6. *Banisch S., Olbrich E.* An Argument Communication Model of Polarization and Ideological Alignment // Journal of Artificial Societies and Social Simulation. – 2021. – Vol. 24. – №. 1. 10.18564/jasss.4434
7. *Banisch S., Shamon H.* Biased Processing and Opinion Polarisation: Experimental Refinement of Argument Communication Theory in the Context of the Energy Debate // Available at SSRN 3895117. – 2021. https://doi.org/10.2139/ssrn.3895117
8. *Betz G.* Natural-Language Multi-Agent Simulations of Argumentative Opinion Dynamics // Journal of Artificial Societies and Social Simulation. – 2022. – Vol. 25. – №. 1. 10.18564/jasss.4725
9. *Mäs M., Flache A.* Differentiation without distancing. Explaining bi-polarization of opinions without negative influence // PloS one. – 2013. – Vol. 8. – №. 11. – P. e74516.
10. *Mäs M., Flache A., Takács K., Jehn K.A.* In the short term we divide, in the long term we unite: Demographic crisscrossing and the effects of faultlines on subgroup polarization // Organization science. – 2013. – Vol. 24. – №. 3. – P. 716-736.
11. *Chkartishvili A.G., Kozitsin I.V., Goiko,V.L., Saifulin E.R.* On an approach to measure the level of polarization of individuals' opinions // 2019 Twelfth International Conference" Management of large-scale system development"(MLSD). – IEEE, 2019. – P. 1-5.
12. *Kozitsin I.V.* Opinion dynamics of online social network users: a micro-level analysis // The Journal of Mathematical Sociology. – 2021. – P. 1-41.

13. *Petrov A.P., Lebedev S.A.* Online political Flashmob: the case of 632305222316434 // Computational mathematics and information technologies. – 2019. – №. 1. – P. 17-28.
14. *Petrov A.P., Proncheva O.G.* Modeling propaganda battle: Decision-making, homophily, and echo chambers // Confer-ence on Artificial Intelligence and Natural Language. – Springer, Cham, 2018. – P. 197-209.
15. *Petrov A.P., Proncheva O.G.* Identifying the Topics of Russian Political Talk Shows // Proceedings of the Conference on Modeling and Analysis of Complex Systems and Processes. – 2020. – P. 22-24.
16. *Kozitsin I.V., Chkartishvili A.G.* Users' Activity in Online Social Networks and the Formation of Echo Chambers // 2020 13th International Conference" Management of large-scale system development" (MLSD). – IEEE, 2020. – P. 1-5.
17. *Schmidt A.L., Zollo F., Del Vicario M., Bessi A., Scala A., Caldarelli G., Quattrociocchi W.* Anatomy of news consumption on Facebook // Proceedings of the National Academy of Sciences. – 2017. – Vol. 114. – №. 12. – P. 3035-3039.
18. *Krueger T., Szwabinski J., Weron T.* Conformity, anticonformity and polarization of opinions: insights from a mathematical model of opinion dynamics // Entropy. – 2017. – Vol. 19. – №. 7. – P. 371.
19. *Albert R., Barabasi A.L.* Statistical mechanics of complex networks // Reviews of Modern Physics. – 2002. – Vol. 3. – №. 1. – P. 47.
20. *Watts D.J., Strogatz S.H.* Collective dynamics of 'small-world' networks // Nature. – 1998. – Vol. 393. – №. 6684. – P. 440-442.