МОДЕЛИРОВАНИЕ РАСПРОСТРАНЕНИЯ ЛОЖНОЙ ИНФОРМАЦИИ ПРИ ИНОКУЛЯЦИИ И ОПРОВЕРЖЕНИИ

Петров А.П.

Институт прикладной математики им. М.В. Келдыша РАН,
Россия, г. Москва, Миусская пл., д.4
petrov.alexander.p@yandex.ru

Аннотация: Предлагается математическая модель процесса распространения ложной информации в социуме при противодействии этому распространению путем как инокуляции, так и опровержения. Инокуляция проводится до, а опровержение - после вброса информации. Проведены численные эксперименты с моделью, результатам дана содержательная трактовка.

Ключевые слова: информационное противоборство, инокуляция, опровержение, математическая модель, численный эксперимент.

Введение

Противодействие ложной информации путем ее опровержения имеет довольно низкую эффективность при отражении информационных атак. Одна из причин состоит в том, что ложная информация распространяется шире, чем ее правдивое опровержение. Это показано, например, в работе [1] на материале распространения примерно 126 тысяч историй в Твиттере. Каждая история была проверена несколькими фактчекерами; на основе проверок истории были классифицированы по трем категориям: истинные, ложные и смешанные (частично истинные, частично ложные). Авторы указанной работы показали, что правдивые истории лишь в исключительных случаях републиковались более чем 1000 пользователей, в то время, как 1% наиболее популярных ложных историй републиковались от 1000 до 100 тыс. раз. Более того, ложные истории распространялись быстрее, т.е. для конкретная численность републикаций достигалась раньше ложными новостями., чем правдивыми. Скорость можно назвать в качестве второй причины низкой эффективности опровержений: они не успевают за опровергаемыми сообщениями. К прочим причинам относятся, например, более высокая стоимость (в терминах времени или иных) поиска истинной информации по сравнению с фабрикацией ложных историй.

В то же, время, проблематика распространения ложной информации становится все более актуальной. Растет не только общественное внимание к ней, но также исследовательский интерес. Характерная статистика приведена в работе [2]: количество научных статей, содержащих точное выражение "fake news" составляло, по данным Google Scholar, лишь 196 в 2003 году, 612 в 2015, и испытало взрывной в последние годы, достигнув 7380 уже в 2017 году. На этом фоне, достигнутое понимание неэффективности опровержений привело к тому, что в качестве оборонительной меры против ложных сообщений все больше рассматривается инокуляция.

Базовая идея инокуляции заимствована из иммунологии. Она состоит в том, чтобы предоставить индивиду ложную информацию в ослабленном виде, тем самым облегчив ему ее отторжение. Наиболее ранние работы по данной тематике относятся еще к 1961 году [3, 4], но взрывное развитие теории происходит в наши дни (см., напр., [5-7] и содержащуюся в этих работах библиографию).

Современные методы инокуляции не только используют идею об ослабленной "той самой" информации, но развивают подход, инокулирующий индивидов от манипуляционных техник. Другими словами, суть новых методов состоит не в том, что "тебе скажут то-то, и это будет ложь", а в том, что "тобой будут манипулировать таким-то способом". Известны реализации этих методов в виде компьютерных игр Bad News [8, 9] и Harmony Square [10], фактически представляющих собой инструменты обучения в игровой форме.

В более широком контексте, данная область исследований относится к тематике информационных противоборств и общественного мнения. Математическому моделированию и эмпирическому исследованию динамики мнений и информационного влияния посвящена обширная литература – см., например, монографию [11] и статьи [12-18].

В настоящей работе предложена математическая модель, описывающая динамику распространения ложной информации при том, что обороняющаяся сторона инокулирует часть социума до вброса ложной информации, и публикует опровержение после вброса. Проведены численные эксперименты, результатам дается содержательная трактовка.

1 Модель

Рассматривается модель с дискретным временем; описываемый ею процесс имеет следующий вид. Атакующая сторона распространяет в социуме численности *N* ложную информацию, против которой обороняющаяся сторона предпринимает две меры: инокуляция и опровержение. Инокуляция охватывает не всех членов социума, а лишь долю νN носит упреждающий характер; она проводится единовременно, ранее момента времени $t=0$. Инокуляция не защищает индивида от ложной информации полностью, она лишь делает его менее восприимчивым. Ложная информация вбрасывается в момент времени $t=0$; в этот момент νN индивидов являются инокулированными, оставшиеся $(1-\nu)N$ относятся к категории восприимчивых. В момент $t=1$ некоторые индивиды уже заражены ложной информацией.

Как восприимчивые, так и инокулированные члены социума могут принять информацию за истинную; в этом случае будем говорить, что происходит заражение, соответствующий индивид переходит в категорию распространителей и начинает распространять информацию другим индивидам (восприимчивым и инокулированным) в виде слуха, т.е. при межличностной коммуникации с ними. С течением времени распространитель может получить опровергающую информацию; приняв ее, он переходит в категорию скептиков. Если опровержение получит восприимчивый либо инокулированный член социума, то он также может стать скептиком. Опровержение распространяется непрерывно вещающими средствами массовой информации с , а также распространяется скептиками. При этом, поскольку правдивая информация распространяется самими индивидами существенно менее активно, чем ложная [19], то в численных экспериментах положено, что соответствующий коэффициент для опровергающей информации на порядок меньше, чем для ложной. Именно, интенсивность распространения ложной и опровергающей информации через межличностную коммуникацию описывается, соответственно, параметрами принято c_1, c_2, и в численных экспериментах принято $c_2 << c_1$.

Положительное принятие информации описывается субъективной достоверностью, которая в отношении ложной информации различается для восприимчивых и инокулированных. Положим, что субъективная достоверность для восприимчивых членов социума имеет вид $\rho_s \exp(-\mu t)$, для инокулированных – $\rho_z(t)\exp(-\mu t)$. Здесь множитель $\exp(-\mu t)$ отражает эмпирически установленное в работе [19] убывание доверия к ложной информации с течением времени после ее опубликования. Другими словами, ложная информация обладает убывающей с течением времени убедительностью так, что, например, восприимчивый индивид, получивший ее на третий день распространения, заразится с большей вероятностью, чем если бы он получил ее на четвертый день (аналогично – для инокулированного).

В указанной работе также установлено, что с течением времени убывает и эффект от инокуляции: различие в уровне доверия между инокулированными и восприимчивыми максимально непосредственно после инокуляции, и со временем убывает до нуля (в условиях проведенного в указанной работе эксперимента, при отсутствии дополнительных поддерживающих инокуляцию мер – примерно за два месяца). Соответственно, введенный выше параметр ρ_s является константой ($0<\rho_s<1$), а функция $\rho_z(t)$ должна обладать следующими свойствами: $\rho_s<\rho_z(t)<1$ в любой момент времени $t\geq 0$ (так как достоверность ложной информации для инокулированных ниже, чем для восприимчивых), $\rho_z(t)$ является возрастающей функцией времени, асимптотически приближаясь к достоверности для восприимчивых ρ_s. Это положение отражает указанное выше свойство убывания эффекта от инокуляции с течением времени. Конкретизируя эти положения, примем $\rho_z(t)=\rho_s(1-E\exp(-\delta t))$, где $E<1$ - эффективность инокуляции; параметр δ описывает скорость убывания инокуляционного эффекта. Достоверность опровергающей информации положим равной для всех групп.; обозначим ее через ρ_2.

Обозначим численность восприимчивых в момент времени *t* через *s*(*t*),. численность инокулированных через *z*(*t*), распространителей и разубежденных – соответственно, $x(t), r(t)$. Положим, что вброс ложной информации позволил ознакомиться с ней доле α как среди восприимчивых, так и инокулированных; тогда численность зараженных ей индивидов из числа

восприимчивых равна $\alpha(1-\nu)N\rho_s$ (соответственно, в категории восприимчивых остались $(1-\nu)N-\alpha(1-\nu)N\rho_s=(1-\nu)N(1-\alpha\rho_s)$), а из числа инокулированных: $\alpha\nu N\rho_z(0)=\alpha\nu N\rho_s(1-E)$ (в категории инокулированных остаются $\nu N-\alpha\nu N\rho_s(1-E)=\nu N\left[1-\alpha\rho_s(1-E)\right]$). Тогда в момент начала распространения слуха имеем

$$s(1)=(1-\nu)N(1-\alpha\rho_s),\ \ z(1)=\nu N\left[1-\alpha\rho_s(1-E)\right],$$
$$x(1)=\alpha(1-\nu)N\rho_s+\alpha\nu N\rho_s(1-E),\ \ r(1)=0. \qquad (1)$$

В соответствии с изложенным выше, динамика описывается следующими уравнениями:

$$s(t)+z(t)+x(t)+r(t)=N, \qquad (2)$$

$$s(t+1)-s(t)=-\rho_s e^{-\mu t}c_1\frac{x(t)}{N}s(t)-\left(b+c_2\frac{r(t)}{N}\right)s(t), \qquad (3)$$

$$z(t+1)-z(t)=-\rho_s\left(1-E\exp(-\delta t)\right)e^{-\mu t}c_1\frac{x(t)}{N}z(t)-\left(b+c_2\frac{r(t)}{N}\right)z(t), \qquad (4)$$

$$x(t+1)-x(t)=\rho_s e^{-\mu t}c_1\frac{x(t)}{N}\left[s(t)+\left(1-E\exp(-\delta t)\right)z(t)\right]-\left(b+c_2\frac{r(t)}{N}\right)x(t), \qquad (5)$$

$$r(t+1)-r(t)=\left(b+c_2\frac{r(t)}{N}\right)\left[x(t)+s(t)+z(t)\right], \qquad (6)$$

Подчеркнем, что параметры c_1,c_2 описывают интенсивность распространения информации, т.е. относятся к распространителям, а множители вида $\rho_s\exp(-\mu t)$ описывают восприятие, те. относятся к "слушателям".

2 Численные эксперименты

Относительно численных экспериментов необходимо сделать следующее предварительное замечание. В связи с дискретностью модельного времени, переменные $s(t),z(t)$, рассчитанные по уравнениям (1)-(6), при некоторых значениях параметров принимают отрицательные значения. Именно, содержательно значимыми являются функции $s(t),z(t)$,убывающие до нуля монотонным образом, однако рассчитанные функции при приближении к значениям $s=0$, $z=0$ иногда принимают отрицательные значения, при сохранении общей тенденции стремления к нулю (условно говоря, дискретность времени приводит к затухающим колебаниям около стационарного решения при том, что решение аналогичного дифференциального уравнения было бы монотонным). Данный эффект является ожидаемым и типичным для систем с дискретным временем. Он не оказывает значимого влияния на общую динамику системы и на содержательные выводы, однако для его устранения были модифицированы уравнения (3), (4): так, вместо уравнения вида $s(t+1)=f(t)$ (где вид функции $f(t)$ соответствует уравнению (3)) в численном алгоритме использовалось уравнение $s(t+1)=\max\{f(t);0\}$. Аналогичная модификация проведена для уравнения (4).

Перейдем к изложению результатов численных экспериментов.

Эксперимент 1. Примем следующие значения параметров:

$$N=10^7\ (10\,млн.\,чел);\ \ b=0,02;\ \ c_1=7;\ \ c_2=0,001;\ \ \delta=0,02;\ \ \mu=0,05;$$
$$\rho_s=0,3;\ \ E=0,4;\ \ \nu=0,2;\ \ \alpha=0,1. \qquad ,$$

График решения показан на Рис.1. Очевидно, при данных значениях параметров инокуляция не оказывает существенной роли в противодействии ложной информации. Численность категории инокулированных, изначально составлявшая изначально 20% (ввиду того, что $\nu=0,2$), убывает до нуля примерно за то же время (5 дней), что и численность восприимчивых. В то же время, численность индивидов, зараженных лозной информацией, за эти же 5 дней возрастает до 90% от

населения, и лишь затем постепенно убывает, причем не ввиду инокуляции, а вследствие разубеждения.

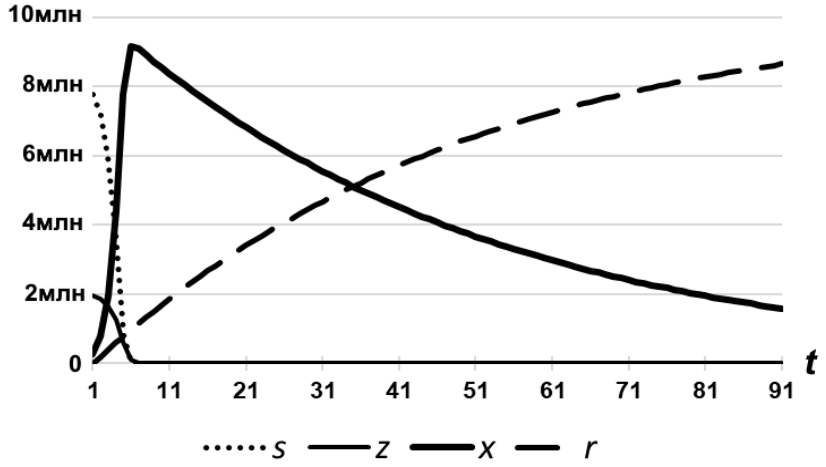

Рис. 1. Эксперимент 1: результаты расчета.

Эксперимент 2. В предыдущем эксперименте инокулированные составляли лишь 20% от общей численности индивидов. Настоящий эксперимент предназначен выявить, увеличится ли существенно значимость инокуляции в противодействии распространению ложной информации, если количество инокулированных будет выше. Соответственно, примем теперь $\nu = 0,7$, оставив значения прочих параметров теми же, что в Эксперименте 1. Таким образом,

$$N = 10^7 \ (10\,млн.\,чел);\ \ b = 0,02;\ \ c_1 = 7;\ \ c_2 = 0,001;\ \ \delta = 0,02;\ \ \mu = 0,05;$$
$$\rho_s = 0,3;\ \ E = 0,4;\ \ \nu = 0,7;\ \ \alpha = 0,1.$$

,

График решения представлен на Рис. 2. Он показывапет, что даже при довольно высокой доле инокулированных индивидов, инокуляция не оказывает существенной роли в противодействии ложной информации. При этом численность зараженных на конец расчета (на 91-ый день) оказывается практически той же, что в Эксперименте 1 (т.е. при доле инокулированных $\nu = 0,2$). Максимальное количество зараженных (достигаемое в день $t = 9$) лишь немного меньше, чем при $\nu = 0,2$.

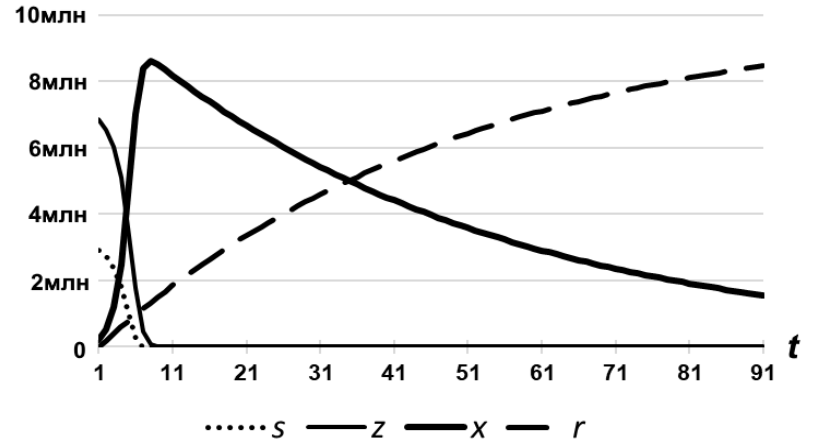

Рис. 2. Эксперимент 2: результаты расчета.

Эксперимент 3. В предыдущих экспериментах эффективность инокуляции принималась равной $E = 0,4$. Другими словами, непосредственно после инокуляции доверие инокулированных к ложной информации составляло 0,6 от доверия тех, кто не был инокулирован. Настоящий эксперимент предназначен выявить, увеличится ли существенно значимость инокуляции в противодействии распространению ложной информации, если эффективность инокуляции будет выше.

Соответственно, примем теперь $E = 0,8$, оставив значения прочих параметров теми же, что в Эксперименте 2. Таким образом, параметры имеют следующие значения:

$$N = 10^7 \ (10\,млн.\,чел);\ b = 0,02;\ c_1 = 7;\ c_2 = 0,001;\ \delta = 0,02;\ \mu = 0,05;$$
$$\rho_s = 0,3;\ E = 0,8;\ \nu = 0,7;\ \alpha = 0,1.$$

График решения представлен на Рис. 3. Он показывает, что и в этом случае инокуляция играет довольно ограниченную роль в противодействии ложной информации. Максимальное количество зараженных заметно меньше, чем при $E = 0,4$ (примерно 7 млн против 8,5 млн в Эксперименте 2), и достигается оно существенно позже (в день $t = 15$). Однако и этот весьма ограниченный эффект действует недолго: так, уже в день $t = 19$ численность зараженных оказывается примерно такой же, как в Эксперименте 2 (6,85 против 6,95).

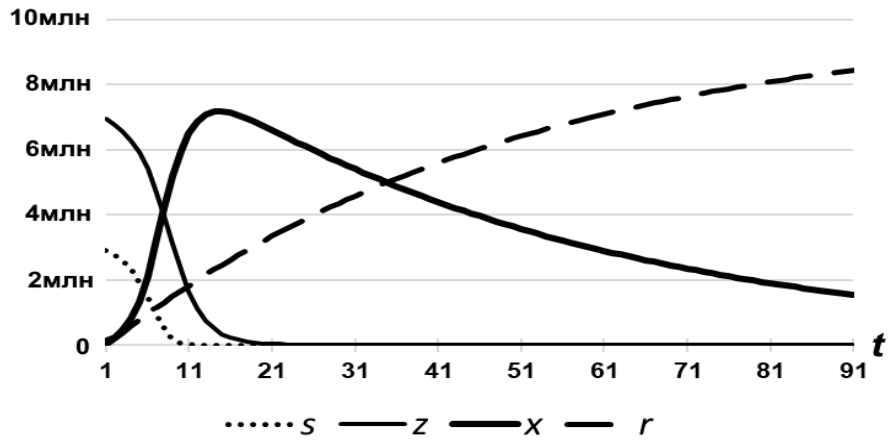

Рис. 3. Эксперимент 3: результаты расчета.

Эксперимент 4. Последним параметром (после ν, E, находившихся в фокусе внимания в Экспериментах 2,3), характеризующим инокуляцию, является скорость убывания инокулирующего эффекта δ. Предположим теперь, что эффект не убывает. Другими словами, если непосредственно после инокуляции доверие инокулированных к ложной информации составляло 0,2 от доверия тех, кто не был инокулирован, то и в каждый последующий момент будет так же. Настоящий эксперимент предназначен выявить, увеличится ли существенно значимость инокуляции в противодействии распространению ложной информации, если выполнено это положение. Соответственно, примем теперь $\delta = 0$, оставив значения прочих параметров теми же, что в Эксперименте 3. Таким образом,

$$N = 10^7 \ (10\,млн.\,чел);\ b = 0,02;\ c_1 = 7;\ c_2 = 0,001;\ \delta = 0;\ \mu = 0,05;$$
$$\rho_s = 0,3;\ E = 0,8;\ \nu = 0,7;\ \alpha = 0,1.$$

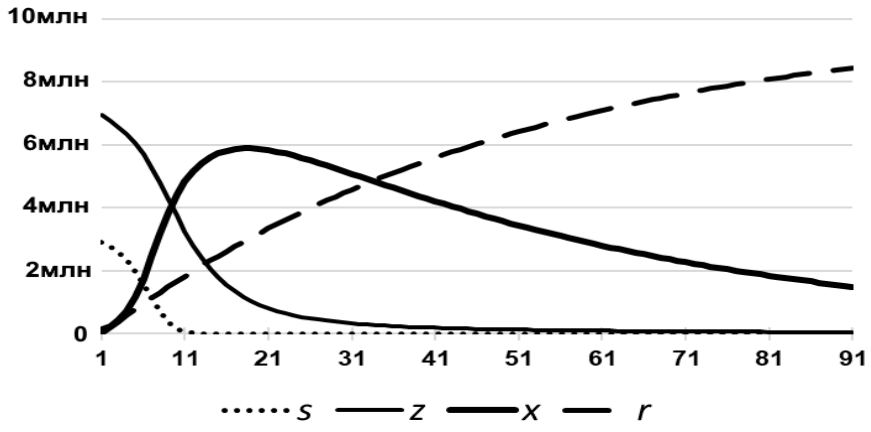

Рис. 4. Эксперимент 4: результаты расчета.

График решения представлен на Рис. 4. Он показывает, что даже при продолжающемся неограниченно долго эффекте инокуляции, она играет довольно ограниченную роль в

противодействии ложной информации. Максимальное количество зараженных 5,9 млн и достигается в день $t = 19$. С течением времени, различие в числе зараженных между случаями $\delta = 0$ (Эксперимент 4) и $\delta = 0,02$ (Эксперимент 3) убывает. Например, при $t = 25$ эти численности равны, соответственно, 5,59 млн. и 6,11 млн., а при $t = 35$: 4,98 и 4,73. Принципиальным является то, что в течение продолжительного времени численность индивидов, зараженных ложной информацией, превышает половину всего населения.

Заключение

Предложена математическая модель процесса распространения ложной информации (fake news) в социуме при противодействии этому распространению как путем инокуляции, так и опровержения. Постановка вопроса связана с тем, что борьба с ложной информацией путем ее опровержения является малоэффективной, так как опровержения не только запаздывают, но также менее широко распространяются. Поэтому приобретает актуальность разработка и изучение превентивных мер, препятствующих распространению ложной информации еще до ее распространения. Под инокуляцией в широком смысле понимается любая мера подобного рода.

В настоящей работе рассматривается разовая инокуляция; относительно нее предполагается, что она затрагивает лишь часть индивидов (а не все население), в определенной мере уменьшает доверие индивидов к ложной информации (но не полностью защищает их от заражения), а также что ее эффект убывает с течением времени. Эти свойства соответствуют эмпирически установленным свойствам информационной инокуляции, проводимой по современным методикам. Численные эксперименты показали, что даже если во всех этих трех аспект параметры инокуляции являются "сильными", она оказывается неспособной создать у общества "коллективный иммунитет" к ложной информации. Основную роль в противодействии fake news играет опровержение; при этом в течение значительного периода времени ложную информацию поддерживает около половины населения или более. Данные результаты получены в предположении, что fake news обладают сравнительно высокой вирусностью. Эта вирусность и является конкурентным преимуществом ложной информации. Эмпирические данные говорят о том, что правдивая информация (опровержение) распространяется индивидами менее интенсивно, чем ложная. В то же время, свойство быть инокулированным вообще не передается от индивида к индивиду, что отрицательно сказывается на эффективности стратегии борьбы с fake news, основанной на инокуляции.

Литература

1. Vosoughi S., Roy D., Aral S. The spread of true and false news online // Science. Vol. 359. 2018. – P. 1146 –1151.
2. *Petrov A., Proncheva O.* (2018) Modeling Propaganda Battle: Decision-Making, Homophily, and Echo Chambers // Artificial Intelligence and Natural Language. AINL 2018. Communications in Computer and Information Science. Vol 930. 2018. Springer. P. 197-209.
3. McGuire W.J., Papageorgis D. The relative efficacy of various types of prior belief-defense in producing immunity against persuasion // Journal of Abnormal and Social Psychology. Vol. 62. 1961. – P.327–337.
4. *Papageorgis D., McGuire W.J.* The generality of immunity to persuasion produced by pre-exposure to weakened counterarguments. Journal of Abnormal and Social Psychology. Vol. 62. 1961. - P. 475–481.
5. *Compton J.* Inoculation theory. The Sage handbook of persuasion: Developments in theory and practice, Vol. 2. 2013. P .220-237.
6. *Banas J.A., Richards A.S.* Apprehension or motivation to defend attitudes? Exploring the underlying threat mechanism in inoculation-induced resistance to persuasion // Communication Monographs, 84(2), 2017. P.164-178.
7. *Compton J., van der Linden S., Cook J., Basol M*. Inoculation theory in the post-truth era: Extant findings and new frontiers for contested science, misinformation, and conspiracy theories // Social and Personality Psychology Compass, e12602. DOI: 10.1111/spc3.12602.
8. *Roozenbeek J., van der Linden S.* The fake news game: Actively inoculating against the risk of misinformation // Journal of risk research. Vol. 22(5). 2019. P. 570–580.
9. *Roozenbeek J., van der Linden S*. Fake news game confers psychological resistance against online misinformation. Nature Humanities and Social Sciences Communications, 5(65) 2019 .
10. *Roozenbeek J., van der Linden S.* Breaking Harmony Square: A game that “inoculates” against political misinformation. The Harvard Kennedy School Misinformation Review, 1(8). 2020.
11. *Chkhartishvili A.G. , Gubanov D.A., Novikov D.A.* Social Networks: Models of information influence, control and confrontation. Cham, Switzerland: Springer International Publishing, 2019. – 158 p. DOI: 10.1007/978-3-030-05429-8
12. *Gubanov D., Petrov I.* Multidimensional Model of Opinion Polarization in Social Networks // 2019 Twelfth International Conference "Management of large-scale system development" (MLSD). Moscow, Russia: IEEE, 2019. С. 1-4. DOI: 10.1109/MLSD.2019.8910967 .

13. *Chartishvili 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), Moscow, Russia, 2019, pp. 1-5, doi: 10.1109/MLSD.2019.8911015.
14. *Kozitsin I.V., Marchenko A.M., Goiko V.L., Palkin R.V*. Symmetric Convex Mechanism of Opinion Formation Predicts Directions of Users' Opinions Trajectories // 2019 Twelfth International Conference "Management of large-scale system development" (MLSD), Moscow, Russia, 2019, pp. 1-5, doi: 10.1109/MLSD.2019.8911064.
15. *Boldyreva A., Sobolevskiy O., Alexandrov M., Danilova V.* Creating collections of descriptors of events and processes based on Internet queries // Proc. of 14-th Mexican Intern. Conf. on Artif. Intell. (MICAI-2016), Springer Cham, LNAI, 2016, vol. 10061 (chapter 26), 2016. pp. 303-314, https://doi.org/10.1007/978-3-319-62434-1_26.
16. *Boldyreva A., Alexandrov M., Koshulko O., Sobolevskiy O*. Queries to Internet as a tool for analysis of the regional police work and forecast of the crimes in regions // Proc. of 14-th Mexican Intern. Conf. on Artif. Intell. (MICAI-2016), Springer Cham, LNAI, vol. 10061 (chapter 25), 2016. pp. 290-302.
17. *Akhtyamova L., Alexandrov M., Cardiff J., Koshulko O.* Opinion Mining on Small and Noisy Samples of Health-related Texts. // Advances in Intelligent Systems and Computing III (Proc. of CSIT-2018), Springer, AISC, 2019, vol. 871, p.1-12.
18. *Akhtyamova L., Cardiff J.* LM-Based Word Embeddings Improve Biomedical Named Entity Recognition: A Detailed Analysis. // Bioinformatics and Biomedical Engineering. IWBBIO 2020. Lecture Notes in Computer Science, vol 12108. Springer, Cham, doi: 10.1007/978-3-030-45385-5_56.
19. *Maertens R., Roozenbeek J., Basol M., van der Linden S*. Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied, Vol. 27*(1), 2021. P. 1–16.