Journal of Computational Social Science (2024) 7:741–776
https://doi.org/10.1007/s42001-024-00256-9

RESEARCH ARTICLE

Towards misinformation mitigation on social media: novel user activity representation for modeling societal acceptance

Ahmed Abouzeid[1] · **Ole-Christoffer Granmo**[1] · **Morten Goodwin**[1] · **Christian Webersik**[2]

Received: 30 March 2023 / Accepted: 7 February 2024 / Published online: 22 March 2024

Abstract

Intervention-based mitigation methods have become a common way to fight misinformation on Social Media (SM). However, these methods depend on how information spreads are modeled in a diffusion model. Unfortunately, there are no realistic diffusion models or enough diverse datasets to train diffusion prediction functions. In particular, there is an urgent need for mitigation methods and labeled datasets that capture the mutual temporal incidences of societal bias and societal engagement that drive the spread of misinformation. To that end, this paper proposes a novel representation of users' activity on SM. We further embed these in a knapsack-based mitigation optimization approach. The optimization task is to find ways to mitigate political manipulation by incentivizing users to propagate factual information. We have created PEGYPT, a novel Twitter dataset to train a novel multiplex diffusion model with political bias, societal engagement, and propaganda events. Our approach aligns with recent theoretical findings on the importance of societal acceptance of information spread on SM as proposed by Olan et al. (Inf Syst Front 1–16, 2022). Our empirical results show significant differences from traditional representations, where the latter assume users' exposure to misinformation can be mitigated despite their political bias and societal acceptance. Hence, our work opens venues for more realistic misinformation mitigation.

Keywords Misinformation · Diffusion model · Monte Carlo simulation · Hawkes processes · Learning automaton · Reinforcement learning · Misinformation dataset · Social media · Stochastic optimization

Extended author information available on the last page of the article

Introduction

In the past decade, Farajtabar et al. (2016) introduced a novel intervention-based approach for incentivizing users on social networks to change their activities [2, 3]. In this paper, we focus on applying such intervention to mitigate misinformation[1] on SM with Reinforcement Learning (RL) agents [5–7]. The purpose is to intervene with the network and learn how to incentivize users to propagate factual information, the latter is also known as a truth campaign on SM [8]. In brief, each RL agent monitors a single user i and must persuade those in her friend or follower zone (e.g., users j and k).

Consider the example when a user k is exposed to misinformation because of manipulative influence from user l. To counter the influence user l has on user k, user i would have to exert effort to persuade user k. Therefore, user i should be incentivized to propagate sufficient counteracting information. Conversely, other users with less victimized friends or followers may, to a smaller extent, need such incentivization.

In practice, the capacity to incentivize users may be limited. Then a problem arises if user i misspends the available incentivization budget, e.g., to convince both j and k when only k needs incentivization [6, 9]. To address this problem, fairness-based mitigation techniques [9] were proposed to ensure that all network users receive fair interventions from a limited incentivization budget.

Intervention-based misinformation mitigation approaches usually utilize an information diffusion model [10] to predict the network dynamics and user propagation patterns at a specific discrete time window. Hence, the RL agents can learn about optimal incentivization policies from the simulated environment by the diffusion model [6]. These RL agents, when interacting with the diffusion environment, form a control over the simulated dynamics of users' activities. The control model tries to optimize a loss function by learning optimal incentivization strategies under budget constraints. This task was solved as a multi-agent knapsack optimization problem [5, 9, 11]. However, there are still some research gaps and open questions to obtain a more realistic information diffusion and misinformation mitigation. Therefore, we discuss below some of the main challenges.

(A) Information Diffusion Accuracy Information diffusion models are necessary for intervention-based misinformation mitigation since applying and evaluating multiple intervention strategies on the real social network is infeasible. In a diffusion model, some users' activities are predicted to simulate and mimic the network dynamics. Traditionally, for the problem of online misinformation mitigation, these activities are the users' temporal propagation of misinformation, and normal information [6]. However, that comes with a challenge, as the model would have an inaccurate prediction of the real-world network propagation of these activities.

[1] The term misinformation is sometimes used to refer to all forms of fake news/content. However, in some literature, misinformation is defined as the unintentional spread of false content, while disinformation is the on-purpose spread [4]. In this paper, we refer to all forms of false content, including political propaganda, as misinformation.

Moreover, such critical drawbacks, if occurred, will also affect the veracity of the optimization decisions made by the control model.

(B) Predicting Users' Engagements SM users' engagements occur when users like, comment, or repropagate other users' content. Extending an information diffusion model to simulate more patterns, such as user engagements, is a clear advantage. For instance, that would answer an important and open research question: how do we model the incidence of engagement between those who spread misinformation or victimized by it — and those who would be incentivized to propagate counterinformation? The latter question is important because political bias can cause people not to be interested in engaging with other ideas when incentivized to do so. Consequently, the learned incentives would be meaningless and not represent the actual behaviors on the network. To that end, users' activity representation in diffusion models must be studied wisely. Therefore, which network propagation attributes should be included in such representation becomes a fundamental question to obtain a robust solution for the problem.

(C) Limitations in Available Datasets The currently proposed solutions for mitigating online misinformation [6, 7, 12] suffer from invoking all critical network features in the mitigation and diffusion models. The latter drawback exists because of the limitations in the available datasets [13], from which diffusion models construct the diffusion prediction function as well. That shows how these datasets [14–19] do not reflect on the advances from social science, a field where the problem of societal interaction is significantly relevant. For instance, a recent study illustrates how societal acceptance [1] on social networks can determine the level of effectiveness when introducing factual information. In the latter study, Olan et al. (2022) proposed a conceptual framework of how the concept of societal acceptance explains how communities in SM form societal circles and deny the acceptance of outliers. A societal circle can be defined as a societal bubble on SM [20] where a circle is a group of biased users agreeing on a particular opinion or idea.

Thus, modeling the mitigation over sequential misinformation diffusion needs modeling of temporal activities such as societal circles formulations (e.g., when users agree and engage with particular ideas), incidents of societal bias (e.g., when users propagate something of a particular bias), and misinformation (e.g., when users propagate false information).

Paper contribution

This work addresses challenges **B** and **C** as introduced above, which indirectly contributes to the challenge **A** since more realistic network dynamics representation could lead to more accurate information diffusion. However, to address all challenges related to **A**, additional efforts are needed to significantly reduce simulation errors in the diffusion model. For **B** and **C**, We propose modeling the diffusion of users' engagements, misinformation/normal information, political bias, and societal circles. Hence, our proposed diffusion model is a multiplex diffusion model [21] where multiple interconnected and interdepended diffusion groups interact. Our hypothesis is as follows. What governs users' activity and how their discussions go

is a universe of ideas. Previous studies on the so-called SM filter bubbles [22] can support the latter assumption. These bubbles' associated ideas construct societal circles, where each circle gathers a subset of people inside it by engaging with the concept it represents. Overlapping between circles may exist, but that does not mean no extreme polarization between them could happen simultaneously.

Polarization causes some of these circles to produce misinformation, which persists in such polarization according to how the SM platforms' algorithms are designed [23]. From there, misinformation circulates through these circles with varying degrees of influence. As a result, reducing polarization and misinformation requires weakening the circles that cause or are influenced by misinformation more than others.

In our proposed solution, a harmful circle is weakened when the number of people engaged with its underlying ideology is significantly reduced. Such counts and their variety are obtained from an information diffusion model that predicts temporal activities such as the propagation of authentic content, misinformation, political bias, and societal circle engagements. We highlight this paper's main contributions below while providing open access to both the mitigation control model source code[2] and a novel misinformation dataset.[3]

- We introduce PEGYPT, a novel misinformation dataset with temporal labels on political propaganda, bias, and societal circle formulation dynamics.
- Based on the above dataset, we introduce a novel technique to represent users' activities on social networks with our proposed Multiplex Controlled Multivariate Hawkes Process (MCMHP) diffusion model.
- We propose a novel optimization loss function that takes temporal bias, propaganda, and information from societal circles as part of its domain and guides the control model reward function.
- We extend the recently proposed intervention-based misinformation mitigation algorithm [9] to support scaling up network size through Monte Carlo-based point process simulation with a small sample size. Further, we couple our novel loss function with the algorithm.
- We provide both quantitative and qualitative analysis to show different behavior between a recently introduced misinformation mitigation loss function [9] and our proposed loss function that considers a more convenient domain attributes such as societal bias and societal circles.

Data collection strategy

We collected data from Egyptian Twitter hashtags which discussed the Egyptian presidential election, 2018. The data were extracted using Twitter API between 24^{th} and 27^{th} of March 2018, a few days before and during election voting days. The data

[2] https://github.com/Ahmed-Abouzeid/MMSS_extended.
[3] https://github.com/Ahmed-Abouzeid/PEGYPT.

extraction process was strategically focused on these days of the Egyptian presidential election, providing a unique opportunity to observe patterns of extreme polarization and political manipulation. This period, marked by societal divisions, was influenced by the significant events and challenges Egypt faced in 2011, including a social uprising and the subsequent restoration of voting rights after decades of disenfranchisement.

Paper organization

The rest of this paper is organized as follows. Section 2 provides related work, briefly explaining related technical details. Then, Sect. 3 illustrates the proposed misinformation mitigation loss function, a novel multiplex information diffusion model, a novel simulation technique, and a novel dataset with multiple temporal events. Empirical results, evaluation, and analysis are given in Sect. 4. A brief discussion about our proposed approach and its limitations and concerns is given in Sect. 5. Eventually, Sect. 6 concludes the work and suggests future directions on the topic.

Related work

The problem of misinformation propagation on SM has attracted attention in the past decade. Both technical and philosophical efforts were made to investigate the nature of the problem, its fundamental concepts, and potential solutions. For instance, recent studies investigated the negative impact of misinformation on society and how SM providers are taking action to reduce misinformation propagation [1]. The latter study highlighted the importance of the societal acceptance concept and its association with online content and SM platforms. In that manner, they stated how the social network assembles ideological sub-networks or circles that try to attract people who share similar values and increase the propagation and polarization inside these common circles. Further, these circles clamp down on outsiders who question or oppose these circles' values.

Moreover, psychological inoculation improved resilience against misinformation on SM [24]. The latter approach applied interventions to users to inform them about the manipulation techniques so they could distinguish fake content from authentic one. In the latter study, one of the primary purposes was to focus on reducing misinformation susceptibility rather than stopping it. The latter scenario of mitigation rather than stopping is more realistic since the nature of the technology makes it impossible to stop the propagation of misleading content completely. For example, in political contexts and bubbled online discussions on SM platforms, the confirmation bias makes people believe in what is aligned with their political beliefs no matter how authentic it is [25].

As a proposed technique for a wide range of tasks, Artificial Intelligence (AI) was utilized [26] to address the problem of online misinformation. There are two tasks where AI can be utilized for the problem. On the one hand, it is the

classification of misinformation, and on the other hand, it is the mitigation of misinformation exposure and its influence on SM users.

There are different approaches being adopted for the misinformation classification task. For instance, content-based [27] Machine Learning classifier approach focused on extracting the textual features of online circulated news articles and their headlines. In the latter approach, word embedding techniques were adopted to represent the semantics of the article's contents. In addition, these features could be derived from visual information like typical images, comics, or deceptive pictures. Such multi-modal approaches took advantage of the combination between text and image-based features and showed more efficient detection for some applications [28].

Further, fake news detection based on contextual information was widely adopted in the literature [29]. In the latter, the content representations considered the co-occurrence between a word i and the context word j instead of only relating words to a whole article or content. Additionally, the social context was modeled by connecting publisher-news relations and user-news interactions simultaneously [30]. The latter technique improved the detection performance in some applications as well.

Despite the significant enhancements from the above-mentioned efforts, different challenges [27] stand against fake news detection. For instance, detecting unseen events became an obstacle since news events would have unseen features during the training of the original classifiers. Furthermore, noisy multi-modality is possible since fusion mechanisms would generate inefficient representations. More importantly, adopting detection approaches is essential, but more is needed — because judging online content or users' authenticity would violate freedom of speech [31]. That is, it became challenging in political contexts to draw a sharp line between what is fake and what is not. Hence, more democratic approaches were needed and proposed as we discuss below.

Recent utilization of RL methods on the problem of online misinformation showed that learned policies that expose social network users to factual information would significantly mitigate the effect of misinformation [5–7, 9]. The mitigation approach can be considered an extension of the detection approach since its first task is to learn users' activity patterns from the classified historical events on an SM platform. A common mitigation technique is truth campaigning [9] where the purpose is to learn an optimal mitigation strategy that incentivizes network users to ensure the optimal delivery of factual information to everybody on the network.

There are different proposed incentivization techniques. For example, the latter can be delivering personalized verified news articles to suit users' reading preferences [32]. However, in RL-based truth campaign, the typical way of incentivization is to learn about the amount of incentives per user that would acquire her to accept propagating the verified information on the network [5, 6, 9]. In the latter approach, based on these optimal incentives, the mitigation model ensures a maximal delivery of authentic content which would achieve maximal mitigation.

The utilization of the RL framework means that an intervention with the network users is conducted. The intervention procedure allows the RL agents to learn

about the user's activity. These activities are simulated with an information diffusion model, commonly a Hawkes Process (HP) [33].

The temporal activities of users on SM are usually logged with annotations in datasets that are used to train an information diffusion prediction function [5, 6]. The latter function predicts the information type it was trained on, e.g., misinformation or authentic content activities. Unfortunately, the available datasets [14–19] need more enriched users' activity information. For example, modeling users' activity only through their dissemination patterns of either true or false content does not inform the diffusion model about other aspects, such as political bias and societal engagement. Hence, this paper proposes a novel representation of users' activity, where temporal patterns of bias, societal engagements, and content authenticity were considered when modeling the RL agents' interventions.

Methodology

This section gives a detailed introduction to our novel users' activity representation and how to utilize that for a solution of misinformation mitigation on SM. Hence, Sect. 3.1 demonstrates our proposed users' activity dataset's collection and annotation processes. Further, Sect. 3.2 explains our proposed MCMHP architecture that encapsulates these representations in a more realistic information diffusion model for how misinformation and other interconnected events circulate over the social network. Sections 3.3 and 3.4 respectively illustrate our novel optimization loss function and our proposed idea of controlling users' activity variables to help optimize the loss, achieving misinformation mitigation and a societal acceptance boost. Eventually, Sect. 3.5 explains our simulation technique for large-scale networks and how we determine the truth campaign incentives for each user.

PEGYPT dataset

The data samples represented three categorized temporal events: political bias, societal circles engagement, and political propaganda. The forms of these events varied between the original tweets, quoted tweets, retweets, and replies from all associated hashtags (check hashtags details here[4]) in the Arabic language. The final numbers for users and events were 10,534 and 36,390, respectively.

Egyptian specialists manually annotated these temporal events while following a systematic approach for automatic verification of labeling consistency in the text or media of a tweet. That was achieved by establishing predefined keywords and some combinations of the latter (check for details here[5]), so when they exist — a particular judgment (label) is made to the content and overwrites the human-given label, if the latter contradicted. Hence, during the annotation process, we aimed to avoid

[4] https://github.com/Ahmed-Abouzeid/PEGYPT/blob/main/tags.txt.

[5] https://github.com/Ahmed-Abouzeid/PEGYPT/blob/main/annotate_propaganda.py.

the human bias factor [34] by these predefined keywords that were agreed upon as a code for either propaganda or political bias. For instance, if the content has religious statements and keywords, it was considered as political propaganda, even if the annotator, despite being religious, did not label it as propaganda. The latter process yielded around 20% of the events' labels to be corrected after some inconsistent manual labeling. We named the dataset as "PEGYPT", abbreviated from the terms: ***P***olarized, ***Egypt***.

In the below sub-sections, we describe the criteria for how we annotated the three categorical events. Further, unlike the limited and static labels in the existing datasets [14–19], we highlight how our novel approach for temporal labeling of social network events opens the venue for an extended analytical capacity of the information diffusion and mitigation tasks, as explained later in Sect. 4. For example, when labeling temporal changes of political bias and societal engagements, we could trace how these variables evolve during an intervention-based truth campaign. In the below also, we provide the statistical properties of the collected social network data to give a better understanding of the context for our experiments.

Temporal bias label

The temporal bias label had three possible values that evaluated whether the user-created or engaged-with content was neutral (0), biased towards (1) or against (-1) the election process. Unlike assuming a static political bias for users [12], our temporal bias captures whether users changed their opinions when they engaged with or generated content over time and hence — had contradicted bias between different content. That helped to trace the changes in the frequencies of bias levels during the conduction of the truth campaign and the misinformation mitigation. The latter guided learning more realistic incentives based on traced users' willingness rather than assuming they would accept whatever incentives they would be offered.

Temporal propaganda label

The temporal propaganda label illustrated the temporal patterns of users with regard to sharing politically manipulative content, in the following sense. The label had two possible values, which described whether the content was political propaganda (1) or not (0). The criteria for the latter were based on whether a user misled readers by using religious expressions or engaging in misleading propaganda to manipulate the facts.

Temporal societal circle label

We defined societal circles as the finite set of different ideologies in a particular online content, where each user engages with one or more ideological circles over time. In that sense, an engaged user of a content ideology means a user who created, quoted, retweeted, liked, mentioned, or replied to that content. The PEGYPT dataset had six societal circles, where a circle ideology in content was defined according to

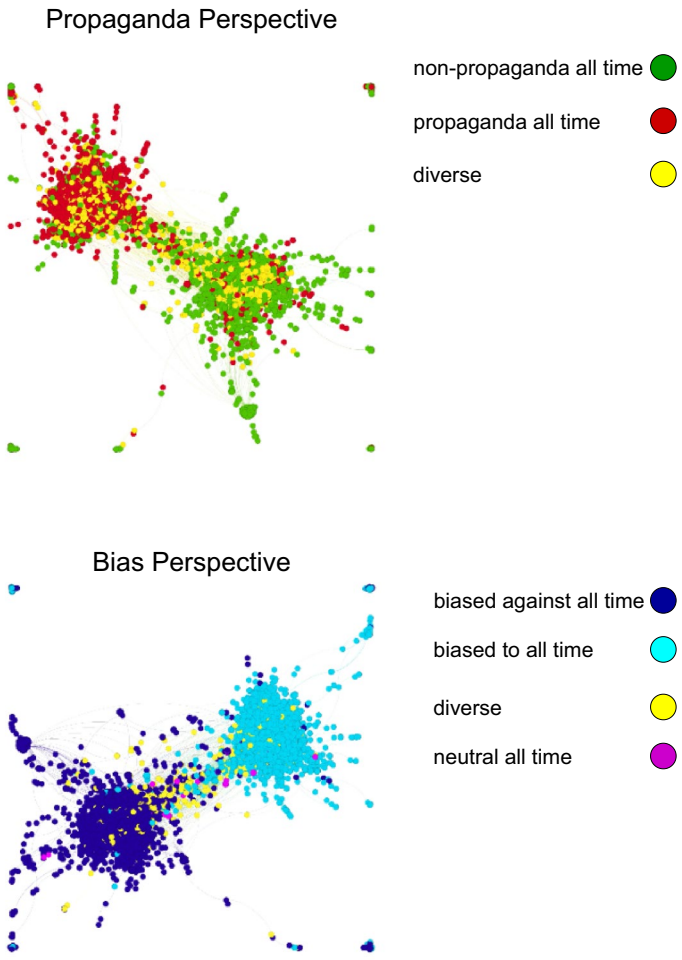

Fig. 1 Colored graph from the PEGYPT network dataset, where nodes and edges represent users and their engagement, respectively. Colors represent the propagation over time of a particular content type

the combination of bias and propaganda labels values. Figures 1 and 2 give a better idea of how these combinations constructed the circles.

We extracted the temporal circle information from each content to obtain the temporal incidents of societal circles events (i.e., temporal labels). In that manner, for each content, we extracted the ideology of the content and associated it with the content creation time. On the same content, we further extracted other engagement forms with their creation times and associated ideologies. For example, when a user generates a primary tweet with a particular bias and authenticity

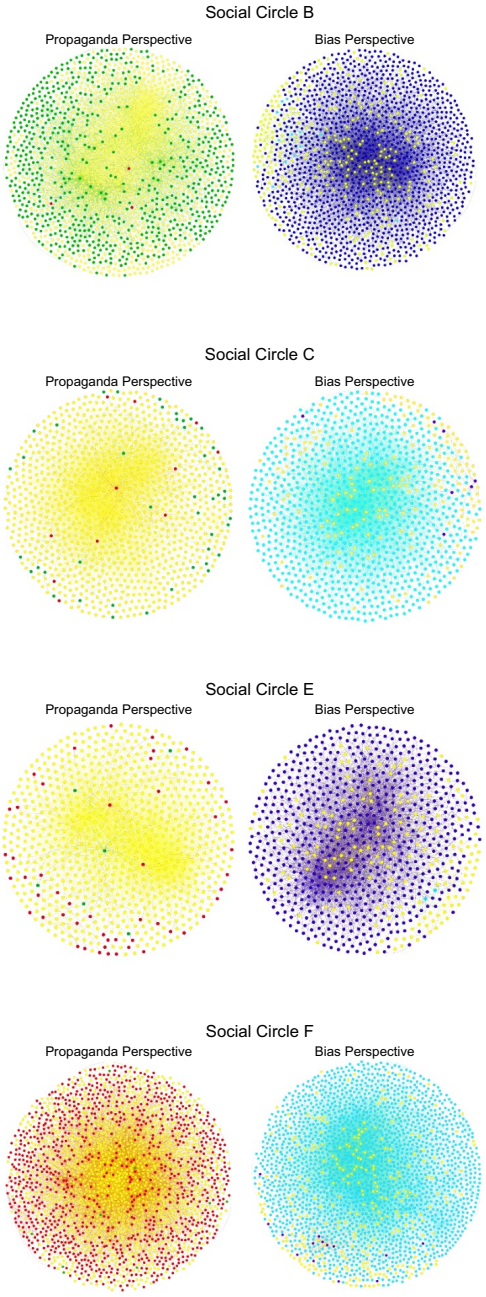

Fig. 2 Extention to Fig. 1: Colored graph from some PEGYPT sub-networks which represent most populated societal circles. Nodes and edges represent users and their engagement, respectively. Colors represent the propagation over time of a particular content type

Table 1 PEGYPT dataset statistics

Metric	Value
Total population	10,534
Number of events	36,390
Number of societal circles	6
Number of graph edges	22,058
Graph modularity	0.596
propaganda users(%)	0.336
Non-propaganda users (%)	0.471
Variant propagation users (%)	0.193
Only biased to-users (%)	0.435
Only biased against-users (%)	0.513
Only neutrally biased-users (%)	0.009
Variant bias users (%)	0.043
Propaganda events (%)	0.536
Non-propaganda events (%)	0.463
Biased to-events (%)	0.515
Biased against-events (%)	0.478
Neutral bias-events (%)	0.006
Biased to + propaganda events (%)	0.429
Biased to + non-propaganda events(%)	0.087
Biased against + propaganda events (%)	0.107
Biased against + non-propaganda events (%)	0.371

level, we consider that as a particular temporal societal circle event. If the latter content had other engagements such as replies, mentions, likes, retweets, or quotes, we consider further temporal events for that societal circle accordingly and label that with its associated engaged user and time of engagement. Thus, a societal circle becomes a structure that changes its density through time (i.e., the number of users represented in its predefined ideology changes over time).

Modeling the temporal changes of societal circles structure is essential to characterize a wide range of the network's user activities. That is why a user's engagement with a societal circle was defined according to whether that user liked, replied, retweeted, quoted, or was mentioned in a tweet belonging to that particular circle concept. That means, engaging with a societal circle did not necessarily mean agreeing with its underlying idea. The reason behind that approach is that we wanted to trace users' exposures to online content realistically, and timestamped users' interaction was the tangible measure we could have found. That was different from previous misinformation mitigation methods [6, 9, 12] where the exposure measures were considered based on the network connections (e.g., following relationships), regardless of whether an interaction will occur. Hence, our approach significantly impacts how the mitigation incentives could be decided since the mitigation algorithm highly depends on content exposure calculation and will be learning from unrealistic estimation if the latter is not appropriately modeled.

Table 2 Example keyword(s) for the "Is-Propaganda=1" label

Keyword(s)	Translation
حق الشهداء	For the sake of maryters
حرب اهلية	Civilian war
تتحول لسوريا	To become like Syria
تبقي اد الدنيا	To become superior over all the world
خونة	Betrayals
عميل لأمريكا	American agent
الله تعالى يقول	God says
ناشط عميل	A betrayal political activist
كلام الله	Words of god

Table 3 Example keyword(s) for the bias label

Keyword(s)	Translation	Label
انزلوا يا مصريين	We Egyptians must go and vote	1
الانتخابات فرحتنا	This election is our joy	1
اختار رئيسك	Choose your president	1
نازلين نكمل المشوار	We will vote to continue the way	1
الزم بيتك	Stay home	-1
بلحة	A sarcastic title people gave on the presidential only candidate	-1

Dataset details

Tables 1, 2, and 3 show some statistics of the PEGYPT dataset and some example criteria keywords for determining propaganda and bias labels, respectively. According to Table 2, some keywords were observed in the collected dataset samples and indicated political manipulation and propaganda. These keywords could also be related to different bias directions since manipulation on the network was from both sides. The made-available dataset files provide the complete details of the associated hashtags and all criteria keywords for both propaganda and bias labels.

To show how misinformation manifiested in the collected social network, Fig. 1 shows colored graphs from the two perspectives: political bias and political propaganda, where nodes colors represent how individual users circulated their contents. Figure 2 also demonstrates an example of the same perspectives for the most crowded societal circles, where some circles were more harmful than others. The complete details of all societal circles and their ideological concepts can be viewed in Table 4. Hence, we can observe how the societal circle F was the most harmful to the top population. It is important to highlight that the data reported in Table 4 does not represent all the population in the dataset, since users were sampled based on top engaging and active ones.

Table 4 Societal circles concepts and population

Circle	Concept	Population
A	Neutral bias + non-propaganda	86
B	Bias against + non-propaganda	1,581
C	Bias towards + non-propaganda	992
D	Neutral bias + propaganda	2
E	Bias against + propaganda	775
F	Bias towards + propaganda	2,084

According to Amnesty's reports on Egypt's human rights situation and witnesses about how fake the election process was[6], we evaluated our mitigation model for the scenario of breaking circle F by incentivizing its members to join other unharmed circles, such as circle B, which was a non-propaganda circle that opposed the election. Thus, in our experiments, we considered the mitigation campaign to oppose the election itself. Further, the mitigation campaign must do that without spreading propaganda to manipulate the public. Hence, a circle with the same bias as our mitigation campaign such as circle E was also considered harmful because it is a propaganda circle.

Information diffusion models

To facilitate an intervention environment for the RL agents to learn about users' activity, an information diffusion model is required to simulate the dynamics of social networks. We simulated the latter using a Multivariate Hawkes Process (MHP). An MHP is a multivariate point process [35] that models the occurrence of temporal or spatiotemporal asynchronous events by capturing the self-and/or mutual excitation (dependencies) between these events. In our context, the MHP is multivariate over the network users.

Through users' activity across the temporal information collected from the PEGYPT dataset, each user was represented by a multiplex HP [21] to predict her future activity on different diffusion groups. In that manner, the diffusion groups represented the temporal patterns of propaganda, non-propaganda, bias towards, bias against, neutral bias, and, eventually, all societal circles' engagement events. Therefore, for each diffusion group, there was a MHP for all users and the relevant group events from PEGYPT data were used to train the diffusion group prediction function over its users.

The associated user HPs are volume-based diffusion models which predicted random counts for all event categories, after being trained on some activity observations in the past. These counts indicated the intensity of the process at a specific time of realization. Hence, an HP for user i can be defined for any diffusion group

[6] https://www.amnesty.org/en/latest/news/2018/01/egypt-authorities-must-cease-interference-in-upcoming-election-and-set-guarantees-for-free-candidacy/.

with its conditional intensity function λ_i. The intensity function has two main components, base intensity μ_i, and an exponential decay kernel function g over an adjacency matrix A. The formal explanation of the conditional intensity function is given by Eq. 1.

$$\lambda_i(t_r|H^{t_r}) := \mu_i + \sum_{t_{s<t_r}} g(t_r - t_s). \tag{1}$$

where μ_i represents a base intensity that models some external motivation to propagate some content. g is some kernel function over the observed history H^{t_r} associated with user i from the discrete-time realization t_s prior to time t_r. g is concerned with the history of some influence matrix A, where $A_{ij} > 0$ if there was an inferred influence between user i and user j, and $A_{ij} = 0$ if not. We utilized an exponential decay kernel function $g = A_{i.}e^{-wt}$, where w is the decay factor where $1 > w > 0$, and represents the rate for how the influence decays over time. For all users, the base intensity vector μ and the influence matrix A were estimated using the maximum likelihood algorithm for the HP [36].

To model the intervention-based mitigation across all diffusion groups, an MCMHP was created, where different diffusion groups were controlled and predicted by a network of Learning Automata (LAs)[9] and MHPs, respectively. We discuss the details of the LAs network in Sect. 3.3.

The diffusion groups encapsulated our proposed novel representation of users' activity on SM. They characterized the interdependence between information veracity-related events, societal bias levels, and societal engagements-related events. Analogically to current approaches of misinformation mitigation [6, 7, 12], Fig. 3 shows our design of diffusion and control models interaction, compared to the typical existing design as shown in Fig. 4.

To evaluate the MHPs predictions, we compared the predicted counts for a diffusion group for all users with the real counts on a test dataset. Therefore, and as shown in Eq. 2, an absolute average error ϵ was calculated to measure how close to reality a MHP prediction was. Where n is the number of users and $N^{\mathcal{H}}$, $N^{\mathcal{R}}$ represents the counts of the arrived events from Hawkes prediction and real data, respectively. The calculation was made between the time stages t_s and $t_s + \Delta$.

$$\mathcal{E}_{t_s+\Delta} = \frac{1}{n}\sum_{i=1}^{n} |[N_i^{\mathcal{H}}(t_s + \Delta) - N_i^{\mathcal{H}}(t_s)] - [N_i^{\mathcal{R}}(t_s + \Delta) - N_i^{\mathcal{R}}(t_s)]| \tag{2}$$

Figure 5 demonstrates how we organized the temporal diffusion group's associated samples from the PEGYPT dataset — to train (estimate μ and A) the MHP where the temporal events counts per user were aggregated into ordered discrete time realizations.

The core idea behind an MHP-based mitigation task is to intensify a particular event in a diffusion group to produce more occurrences against another harmful event type(s). Users-associated HPs for the to-be-intensified event category should be modified to achieve that. Hence, let s_i be the incentivization amount decided for user i, and the modified HP for mitigation purposes can be redefined by Eq. 3.

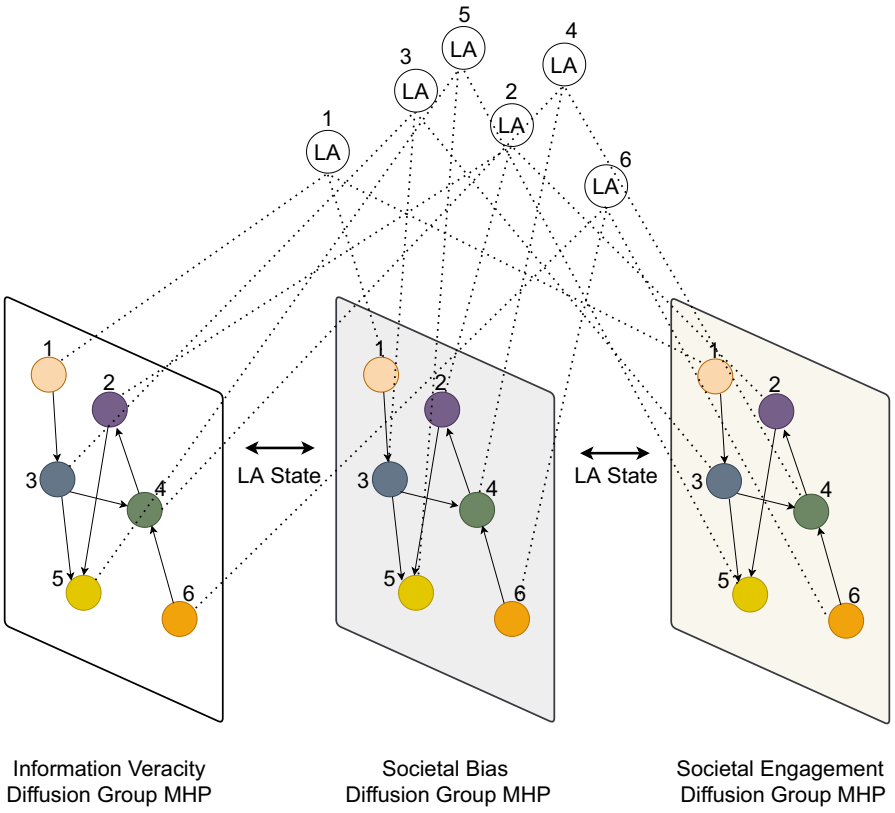

Fig. 3 A toy example of a social network with 6 users and the proposed design of MCMHP interaction, where each LA state is shared between all diffusion groups

$$\lambda_i(t_r|H^{t_r}) := s_i + \mu_i + \sum_{t_s<t_r} g(t_r - t_s). \tag{3}$$

where s_i is a parameter to be optimized for user i for the mitigation targeted event category, the optimal values for s across all users were governed through a restricted incentivization budget knapsack optimization and a loss function to dictate the rewarding of a RL agent, associated with each user. In our solution, we associated an LA for each user as the individual RL agent that learns the optimal value of s_i.

Controlling of multiplex diffusion groups

As a control model over the stochastic MHP environment, we utilized the LA [37] for its easy decentralized implementation and lightweight computation when compared to traditional RL techniques adopted for the probelm of misinformation mitigation [6, 7]. The LA learns by interacting with the MCMHP and updates its actions or state transitions according to the stochastic signal from a MHP counts-based loss

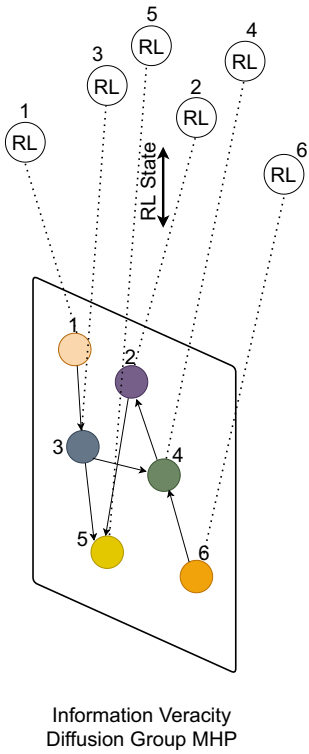

Fig. 4 A toy example of a social network with 6 users and the typical design of MHP interaction with a control model

function. In our proposed solution, each LA is attached to each user to learn an optimum/sub-optimum state s^*, where the latter represents a discrete decision value for each user's incentive in the mitigation campaign. As indicated in Fig. 3, these incentives (LAs states) are shared across all diffusion groups to embrace the interdependencies between the different aspects of users' activity. The LA seeks convergence at such an incentive value by optimizing the latter through the loss function. The latter dictates the potential reward or penalty of the LAs, and the LAs updates their states accordingly. The loss function evaluates its gradient when an LA increases its state and causes new predicted volumes from the different diffusion groups in the MCMHP. Hence, if the loss slop declined, then the LA should be rewarded. If inclined, the LA should be penalized.

Figure 6 demonstrates how challenging optimizing such loss function through each associated LA state transition, where optimal states could be non-stationary due to the interdependencies and complexity between all diffusion groups, i.e., some users' optimum value s^* will determine the optimum s^* for others. The latter property persists due to the mutual dependencies between users on the MHP-generated dynamics. That means how many incentives a user i would need could make it unnecessary for an engaged user with user i to have many incentives when both share the same with other engaging users. Therefore, for user i, if s_i is optimum and consequently, for user j, s_j is not. Then, given $s_i^{'}$ as another possible incentive value for user i, $s_i^{'}$ could still be optimum while s_j is also optimum. This non-stationarity

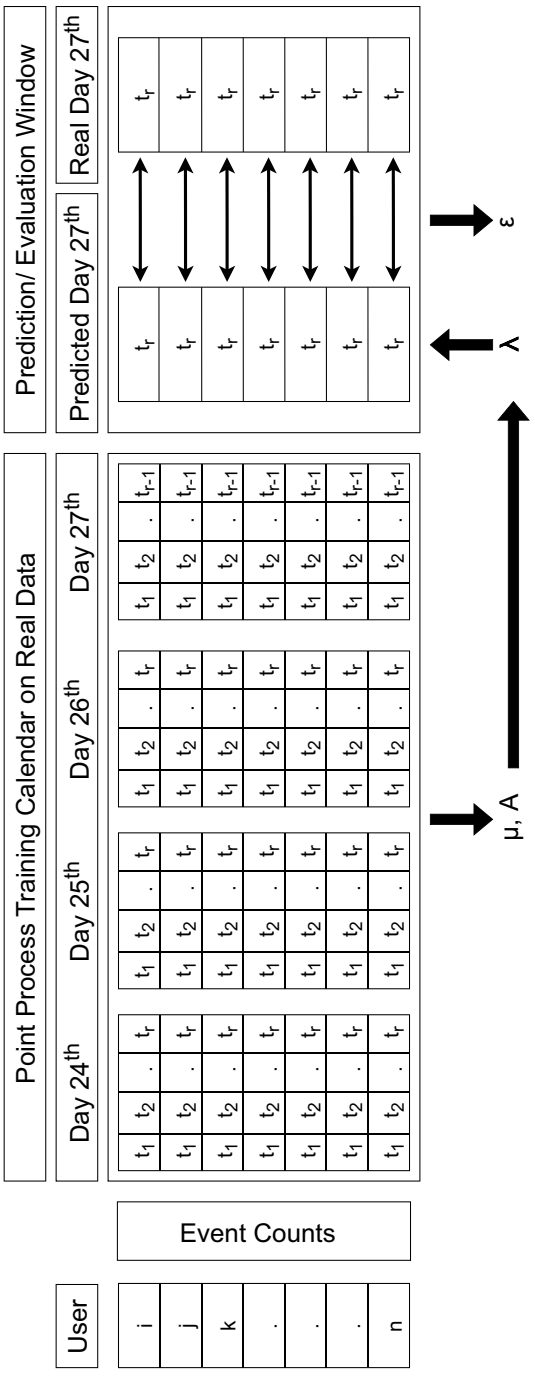

Fig. 5 Feeding the MHP with a diffusion group's samples from the PEGYPT dataset

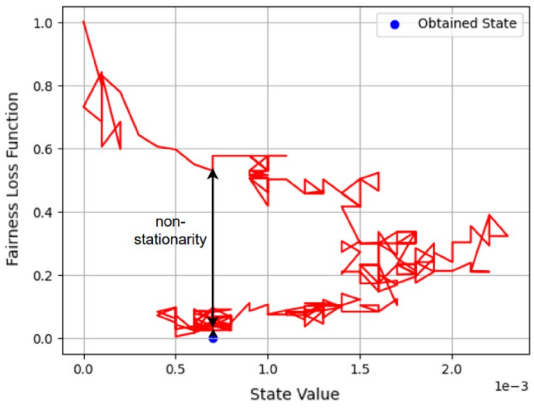

Fig. 6 Non-stationarity of an optimal automaton state in its individual loss function trajectory over time

occurs then because the LAs while consuming the incentivization budget — have to intervene with their associated users in sequential order, not simultaneously.

When saddle points occur, particular LA state transitions and rewarding techniques are applied as proposed by Abouzeid et al. (2022) [9] since we utilized the same LAs network-based control model. The complete details of how each LA learns its incentive value and updates its state transition probabilities are given in Appendix A.

We extended the LAs network-based control model with a Monte Carlo simulation technique [38] over multiple interventions $\{e^1, e^2, ..., e^*\}$. Hence, the control model could interact with random samples instead of the whole network. We repeated the latter procedure over hundreds of sampled networks, and then we calculated the expected values of individual converged s^* values over all samples. We believe this approach opens the venue for scaled-up misinformation mitigation frameworks where the real network size would reach hundreds of thousands of users and more. Moreover, obtaining a probability distribution for a user's decided incentive allowed us to measure the level of uncertainty in the solution (see sect. 4.3).

Optimizing societal acceptance with fairness

The criteria for successful mitigation were based on how eventually the sampled network users would be less exposed to the harmful content since the incentivization should boost the amount of authentic content on the network. Therefore, the optimization task was to reduce a total loss function during the intervention. To achieve the latter, an LA per user conducted the intervention by suggesting a shared incentive value to modify the associated HP diffusion group by which the political manipulation would be mitigated. Thus, we wanted to incentivize the diffusion groups' events of non-propaganda, bias against, and societal circle B with the same shared incentive value.

During learning, after an intervention step e, a dedicated individual loss function is responsible for the evaluation of the current incentive of its user. At the same time, the MCMHP predicted temporal events information was given as the

function domain. Hence, for an individual user i, all other users' predicted activities (from all diffusion groups) were passed to the individual loss function i. Ideally, the total loss function should converge to a steady point after multiple interventions across all users.

We extended the fair mitigation loss function introduced by Abouzeid et al. (2022) [9]. In the latter, the distribution of incentives was conducted according to user needs. In that manner, we keep maintaining the concept of fair incentivization. However, in addition to representing only temporal events for misinformation and authentic content, we propose additional information on the temporal societal circles and temporal bias, to model the occurrence of engagement and its nature, respectively. That means we predict the propagation of authentic content (e.g., non-propaganda), propaganda, engagement of users with societal circles, and eventually, the bias directions of users at a specific intervention step e and time realization t_r. We think such a combination gives more close-to-reality dynamics from diffusion modeling and characterizes the societal acceptance concept that governs social networks [1]. Equation 4 and Eq. 5 demonstrate our novel loss function.

$$\min \mathcal{F}(s_U) := \sum_{i=1}^{N} \Lambda_i'(s_i) + \mathcal{F}(s_i), \text{ where } \mathcal{F}(s_i) := \sum_{j=0}^{n} (2 - R_j^{s_i} - \Lambda_j(s_i))^2, \quad (4)$$

$$\text{subject to } \sum_{i=1}^{|s_U|} s_i, \text{ where } s_i \in [0, C]. \quad (5)$$

The term s_U represents the set of passed users' incentives, where the sum of the latter set cannot exceed the incentivization budget C as demonstrated in Eq. 5. Further, as indicated in Eq. 4, an individual loss function for a user i is evaluated first by measuring how the incentive value s_i affected all other users with an engagement relationship to i. Hence, the term $R_j^{s_i}$ defines the exposure counts ratios between non-propaganda npg and propaganda pg events for all users that can engage with user i (e.g., her followers) at a particular time realization t_r.

A user i exposure to a particular event category (e.g., non-propaganda) at a particular time realization t_r is the count of all events from users that user i can engage with (e.g., her followee) at t_r. Equation 6, Eq. 7, and Eq. 8 show how $R_j^{s_i}$ is calculated, while a user-associated ratio closer to 1 means a boosted non-propaganda exposure. A ratio that exceeds 1 means an unnecessarily high incentive value assigned to that user, which indicates unfairness according to [9]. In Eq. 6, ξ is a tiny smoothing factor with a value close to 0, to avoid division by Zero when propaganda events do not exist for some users. Also, b is a mitigation balance factor [9] to satisfy a mitigation campaign threshold. For example, $b = 2$ if successful mitigation means the exposure to non-propaganda should be at least twice the exposure to propaganda, and hence the unfairness is perceived if the ratio exceeds 2 not 1. The symbol A indicates the network structure adjacency matrix where $A_{ij} = 1$ if j follows i on the network, and $A_{ij} = 0$ if not.

$$R_i^{t_r}(s_i) := \frac{\xi + npg_i^{t_r}(s_i)}{(\xi + pg_i^{t_r}) \cdot b}. \tag{6}$$

$$pg_i^{t_r} := \sum_{s=0}^{t_r} \sum_{j=1}^{n} A_{ij} \cdot pg_j^{t_s}, \tag{7}$$

$$npg_i^{t_r}(s_i) := \sum_{s=0}^{t_r} \sum_{j=1}^{n} A_{ij} \cdot npg_j^{t_s}(s_i), \tag{8}$$

The term $\Lambda_j(s_i)$ is calculated according to Eq. 9, and it represents a joint probability of two events. First, c_j, which is the probability user j who follows user i — would engage with the societal circle to which the mitigation campaign tries to attract people. Second, the probability that j being in the same bias of the mitigation campaign and is denoted as $bias_j$. It is essential to highlight that such probabilities are calculated after applying the incentives s_i and s_j from the associated interventions, which would change the generated counts for bias and societal engagement HP events. Hence, $\Lambda_j(s_i)$ measured the probability of the societal engagement with the circle we seek acceptance of its concept, and the probability of agreeing with that circle during such engagement.

$$\Lambda_j(s_i) := P(c_j) \cdot P(bias_j) \tag{9}$$

While interventions cause different incentives and accordingly different diffusion volumes, given an increased value of $\Lambda_j(s_i)$ will decrease the loss function, and the associated LA_i will be rewarded.

The individual loss for user i could also be increased by $\Lambda_i^{'}(s_i)$, which represents the probability of user i not being in the same bias direction of the mitigation campaign. That means no matter how engaging users with i would agree and engage with the circle we seek — the loss will always be high if user i's bias disagrees with the mitigation campaign. The latter mechanism means that users will consume incentives wisely and according to their probability of accepting the incentives instead of naively assuming they would. Equation 10 shows how the latter probability is calculated.

$$\Lambda_i^{'}(s_i) := 1 - P(bias_i) \tag{10}$$

Monte Carlo simulation

Let us assume the network sample has n users, where $n = 3$ (see Appendix A), and each associated LA has the state depth M ($M + 1$ possible incentive values). Then, we can demonstrate the following procedure. Let the users i, j, and k be the sampled network users at intervention step e. Then, $s_U = \{s^e{}_i, s^e{}_j, s^e{}_k\}$ are the discrete state values of the associated LAs at e. Hence, the converged states and final obtained

results from the interventions can be assigned to the below-modified HPs diffusion prediction functions for the three given users. The modified HPs should suggest an optimum or sub-optimum predicted activity on the network if the obtained state values were passed as incentives. The latter should satisfy the minimization of the total loss function $\mathcal{F}(s_U^*)$ in Eq. 4. Where $\sum_{i=1}^{|s_U^*|} s_i \leq C$.

Equation 11, Eq. 12, and Eq. 13 together construct the incentivized users on the network for a particular event type in a specific diffusion group. Thus, the MCMHP can be viewed as replicating these incentives for all desired events in targetted diffusion groups. For instance, the optimal value s_i is shared across non-propaganda, bias-against, and circle B engagement events to incentivize their associated HPs. Same concept applies for all users.

$$\lambda_i(t_r|H^{t_r}) := {s^{e^*}}_i + \mu_i + \sum_{t_s<t_r} g(t_r - t_s). \tag{11}$$

$$\lambda_j(t_r|H^{t_r}) := {s^{e^*}}_j + \mu_j + \sum_{t_s<t_r} g(t_r - t_s). \tag{12}$$

$$\lambda_k(t_r|H^{t_r}) := {s^{e^*}}_k + \mu_k + \sum_{t_s<t_r} g(t_r - t_s). \tag{13}$$

Since multiple samples are taken during the Monte Carlo sampling procedure, the final determined value for any s is the expected value of the random variable s on its distribution. Hence, for the user i, given a converged random variable $s_i^{e^*}$ from w Monte Carlo samples, the vector $s_i^* = \{{s^*}_1, {s^*}_2, ..., {s^*}_w\}$ represents an example for the possible obtained values from converged automaton LA_i state over w samples, where the user i was sampled w times. Further, the distribution vector $d_i = \{p({s^*}_1), p({s^*}_2), ..., p({s^*}_w)\}$ represents the probabilities for s_i^* entries. Therefore, the final incentive value for a given user i is the expected value for s_i^* over its samples.

Equation 14 shows how the final incentives were determined, where the final incentivization vector for all users is a vector of all expected values calculated over all sampled networks, where U is the set of all users.

$$s_U^{**} = \{\forall s_i^* \in s_U^* : E[s_i^*] = \sum_{l=1}^{w} s_l^* p(s_l^*)\} \tag{14}$$

Empirical results

Experiment setup

In our experiments, we considered a subset of the PEGYPT dataset where only users with high engagement frequencies were selected. The avoiding of sparsity was necessary for the MHP parameters estimation since the latter requires a sufficient

Table 5 The social network used in the experiments as a subset of PEGYPT

Metric	Value
Number of users	940
Number of events	20,084
Only biased towards-users (%)	0.44
Only biased against-users (%)	0.50
Variant bias users (%)	0.06
Propaganda events (%)	0.55
Non-propaganda events (%)	0.45
Max number of events per user	177
Min number of events per user	6
Number of societal circles	5
Number of graph edges	6,619
Graph modularity	0.519
Graph density	0.015

number of events per user. Furthermore, high engagement was essential to study typical social network dynamics where extreme political polarization and propaganda govern the network. Hence, we extracted users with at least six temporal events while keeping similar percentages of propaganda and bias levels as in the original PEGYPT dataset. The final social network had 940 users and 20, 084 temporal events. Table 5 shows the complete details of the obtained social network for the experiments.

We used a sample size of 100 users to construct the sampled networks during the Monte Carlo simulation. We also ran the sampling 100 times to ensure each user will have a probability distribution of the obtained incentives to calculate its expected value. We utilized a time realization period of 180 minutes for the time realization structure. That means events per user (see Fig. 5) were grouped every three hours and passed to the MHP model. The latter structure helped estimate the MHP parameters as the grouped event counts were enough to infer the influence matrix A and the base intensity μ. We set the Knapsack budget $C = 2$ and LA state depth $M = 500$.

From the final 940 users' network, we established eleven MHPs to model the different behavioral aspects of the network via a multiplex diffusion. We ran the LA control model on two different environments setup based on two utilized loss functions for the optimization. The latter setup allowed us to monitor how our proposed societal acceptance representation constructed another environmental behavior for the LA control while learning the incentives.

To mitigate the misinformation caused by political propaganda, we incentivized a group of MHPs through the shared incentive value being learned. For example, when a user is incentivized to create or retweet non-propaganda content, the same content declares a particular bias direction. The latter, combined with the non-propaganda content, belong to a specific societal circle as introduced earlier in Table 4. Hence, the non-propaganda event category was not the only modified MHP — but

Table 6 MHP simulations performance evaluation with a flag indicating the incentivized MHPs

MHP	Z	$\mathcal{E}$	Incentives
Bias-towards	0.59 ± 0	0.30 ± 1.31	No
Bias-against	0.55 ± 0	0.35 ± 2.32	No
Bias-against-sampled	0.38 ± 0.25	0.37 ± 1.44	Yes
Propaganda-sampled	0.30 ± 0.24	0.52 ± 1.25	No
Non-propaganda-sampled	0.32 ± 0.28	0.48 ± 1.41	Yes
Circle A	0.02 ± 0	0.02 ± 0.26	No
Circle B	0.27 ± 0	0.23 ± 1.47	No
Circle B-sampled	0.39 ± 0.25	0.37 ± 1.50	Yes
Circle C	0.02 ± 0	0.06 ± 0.62	No
Circle E	0.09 ± 0	0.00 ± 0.07	No
Circle F	0.01 ± 0	0.03 ± 0.28	No

the relevant events categories for both bias and the particular societal circle were also intensified with the exact amounts represented by the LA state.

To replicate the results and clarify how each MHP was configured and established, we demonstrate the eleven MHPs simulations in Appendix B with their customized configurations and purposes. Further, the experimental network had only five circles, as indicated in Table 5 since circle D had two members only, and that was challenging to simulate. Nevertheless, that did not influence the validity of our experiments.

MHP simulation evaluation

We ran the eleven simulations and reported their results in Table 6. We adopted two evaluation metrics to measure how each MHP was reliable enough for the prediction. First, we calculated the average absolute difference error $\mathcal{E}$ as explained earlier in Eq. 2. We then applied Z-statistic to compare the predicted counts with the actual counts.

As indicated in Table 6, we obtained lower $\mathcal{E}$ and Z values. For more detailed information about the MHPs simulation performance, see Appendix C.

Control model evaluation

In this section, we evaluated and compared our proposed loss function to the previously introduced mitigation fairness loss function [9]. We refer to our proposed loss as *Societal Acceptance + Fairness* since the latter still holds the fairness concept when distributing the incentives. At the same time, it is essential to highlight that it was not feasible to evaluate other control models [6, 7] since their structure depends on an entirely different dataset and representations, where temporal bias and societal circles were not modeled. Further, this work's main focus was to assess the novel representation of users' activities. Hence, we utilized the same control model

Table 7 Control model obtained performance on utilizing different optimization loss functions

Metric	Fairness	Societal Acceptance + Fairness
Propaganda Mitigation	**0.89**±.05	0.88 ± .05
Polarization Mitigation	0.23 ± .10	**0.26** ±.09
Societal Acceptance Boost	0.16 ± .03	**0.19**±.05

The result is an average over 3 independent runs

proposed in [9] to extend its fairness loss function with the societal acceptance concept. We employed three evaluation metrics as below.

- **Propaganda Mitigation:** a traditional mitigation evaluation metric [5, 9] to calculate the percentage of how much reduction happened on the users' exposure [6] to propaganda through the engagement relationships with each other. Equation 15 illustrates how this metric was calculated, where x and y are the political propaganda percentages after and before mitigation, respectively. Hence, the higher this metric, the better.

$$\text{Propaganda Mitigation} := 1 - \frac{x}{y}, \text{ where } x \leq y : y \neq 0 \tag{15}$$

- **Polarization Mitigation:** evaluated the percentage of how much harmful polarization was mitigated on the network. For instance, since the campaign task was to convince users against the manipulation in the election, that metric measured how the bias-towards the election among users was lessened after optimizing the incentives. To measure that, the probability distribution over the three bias levels of each user was calculated first, and then we calculated an average percentage of the bias-towards over all users. The latter was calculated twice: once when using the learned incentives to calculate the distribution from the MHP predictions and once when there was no intervention at all. Thus, the polarization mitigation was calculated following the same concept as in Eq. 15.
- **Societal Acceptance Boost:** similar to the above metric, but measuring the percentage of how much the societal acceptance increased on the network. We defined societal acceptance as the breaking of circle F by letting its users accept ideas from circle B. That means we measured the joint probabilities of being engaged with circle B and being biased-against the election (see Table 4). Similarly to the above metric, we calculated the probability distributions over circle F members to measure how far the intervention succeeded in breaking circle F and allowing its users to accept the societal circle B concept.

Table 7 shows how our proposed societal acceptance representation outperformed (bold numbers indicate the outperforming method) the traditional fairness-only when mitigating polarization and boosting societal acceptance during the misinformation (i.e., propaganda) mitigation. However, we can observe that the percentage of propaganda exposure mitigation was significantly higher than the percentages in

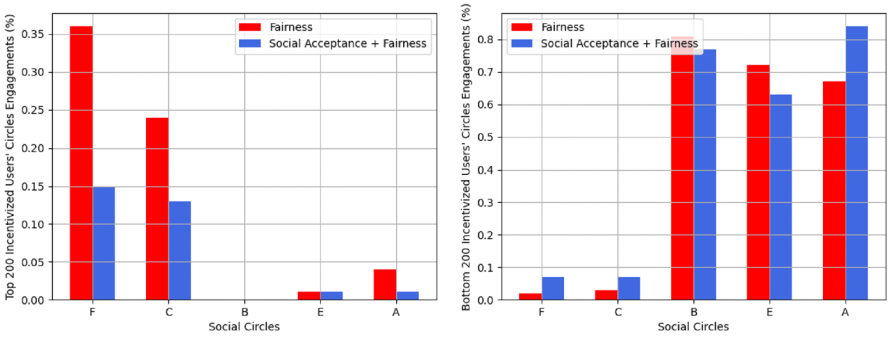

Fig. 7 Incentivized users' circles engagement

both polarization and societal acceptance. We believe that was due to the traditional less strict definition of propaganda content exposure and its mitigation metric [6, 9]. The latter usually consider the counts of events a user is assumed to access through a following/ engaging relationship on the network [6]. However, we believe that would be a naive assumption since following relationships or past engagements do not guarantee actual exposure in the future. Therefore, it was essential to adopt more strict metrics from our proposed representation, such as the actual dynamics of societal acceptance and polarization, which estimated how likely an engagement would occur and to what degree it would be an agreeing engagement inherited from its associated bias. Therefore, our proposed novel representation allowed for calculating the three metrics together, which gave a better justification for the performance.

One of this paper's main motivations and purposes was to analyze the achieved mitigation efficiency to verify what it represented and how the control model learned the incentives. That is because a computational social model evaluation is considered one of the most challenging tasks [39] since the latter lacks a systematic pattern to consider as ground truth. Therefore, we propose the below analysis to help apply some quantitative and qualitative analysis on the model's performance.

Analysis

Figures 7 and 8 demonstrate the difference in behavior between the two mitigation loss functions despite their similar propaganda mitigation performance captured in Table 7. For instance, Fig. 7 on the left side explains how the top 200 incentivized users' engagement was distributed among the different societal circles, i.e., the top most users who consumed the incentivization budget and their societal engagements. We observe that the fairness loss function-based mitigation consumed most of the incentivization budget on users who contributed to around 60% of the engagement in circles F and C. Although circle F was the most harmful circle and circle C also had a different bias than the incentivization campaign (see Table 4). On the contrary, after considering the societal acceptance representation, our proposed loss function consumed most of the budget on less than 30%

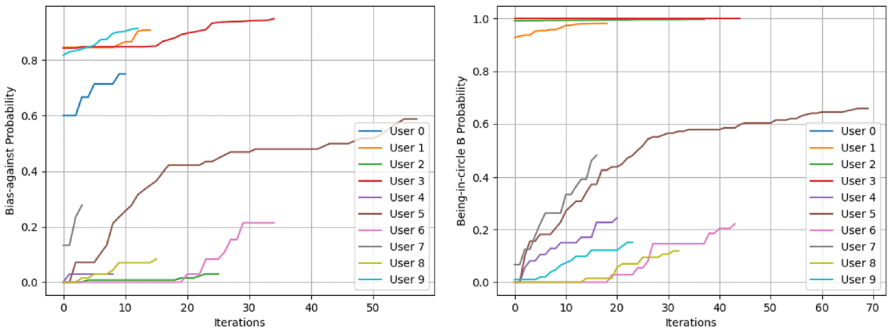

Fig. 8 Example of breaking the societal circles by incentivizing some users to circle B

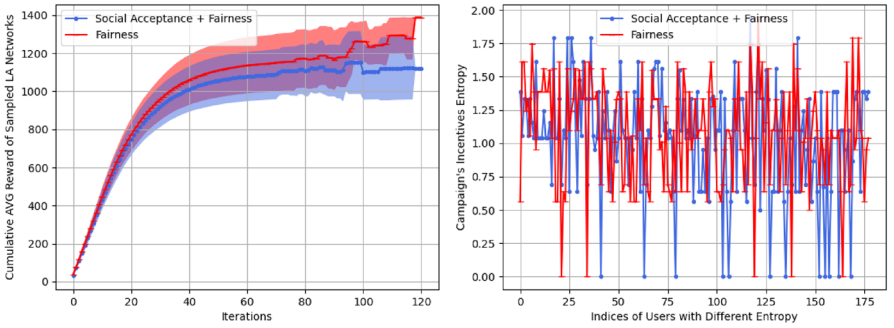

Fig. 9 Average cumulative rewards during incentive learning with entropy of the finally decided incentive value

of these circles' contributors. The latter behavior indicates how the temporal bias and societal circles' information matured the mitigation more and incentivized users based on the probabilities of accepting the incentive while the fairness loss function incentives were given irrationally.

Moreover, on the right side of Fig. 7, lesser distribution of incentives was the other way around for users engaged with the circles F, C. That indicates how vital these circles were for the incentivization campaign to target and assign more incentives. However, that was done more wisely by the societal acceptance loss function.

Further, Fig. 8 gives an example of how we break the other circles to push users to join circle B by engaging with it and accepting its ideology, not engaging by disagreeing. Hence, we observe how the probability of being biased-against and being engaged with circle B increased. The latter represents modeling the engagement occurrence, while the former models the acceptance of that engagement since circle B represents a bias-against concept. That also demonstrates how representing the temporal bias and societal circles' engagements allowed for tracing and analyzing the associated users' activities of these events.

Eventually, on the left side of Fig. 9, we show how the LAs environment, characterized by our societal acceptance representation, was more strict and gave fewer rewards. We believe that was due to the more interdependent variables considered in the societal acceptance-based loss function. However, such rigidity helped achieve higher polarization mitigation and societal acceptance in addition to slightly more certainty of the learned incentives. The latter can be viewed on the right side of Fig. 9, where we calculated the Shannon entropy of the individual incentives' probability distribution which was obtained over the Monte Carlo sampling. We can observe there were more users with significant zero entropies when the societal acceptance loss function was applied. The entropies values in Fig. 9 are only for users with different obtained entropies between the two loss functions.

Discussion

Unlike the recently proposed work [5–7, 9, 12], instead of directly modeling the misinformation volumes and exposures, and learn incentives accordingly, we first model the relevant network dynamics that derive these exposures. The latter extended the analytical capacity of the solution as demonstrated in Fig. 8. However, the reader might wonder about the reason behind not modeling societal acceptance directly instead of modeling the bias and engagement separately. That means defining the temporal societal circles based on acceptance rather than engagement in general. Then, modeling the temporal societal circles' acceptance by an HP to predict the acceptance in the next time realization. In the latter scenario, we will lose the capability to trace and analyze the detailed users' activity, such as the interaction with contents, either by agreeing or disagreeing. The latter information is crucial for any further analysis required on the network.

We extracted user engagements from the direct engagement relationship in the historical data to evaluate the influence of incentives during the intervention. However, indirect engagement or influence could also be considered in future attempts. For instance, if user i engages with user j, and user k engages with user j, then users k and i could be considered indirectly engaging together. This influence-cascading technique could also be applied when we consider other influence patterns instead of engagements, such as the following relationships.

Conclusion and future work

The social sciences are studying the societal acceptance concept on social networks to extract the key elements that can describe human behavior regarding information dissemination. Recent efforts revealed the importance of understanding the relationship between fake news, social network platforms, and societal acceptance. Therefore, this paper considers the interdependencies between the latter in a proposed computational social model for mitigating online misinformation. Our proposed model encapsulates novel representations of users' activity, such as temporal polarization patterns, community engagement, and propaganda dissemination. Derived

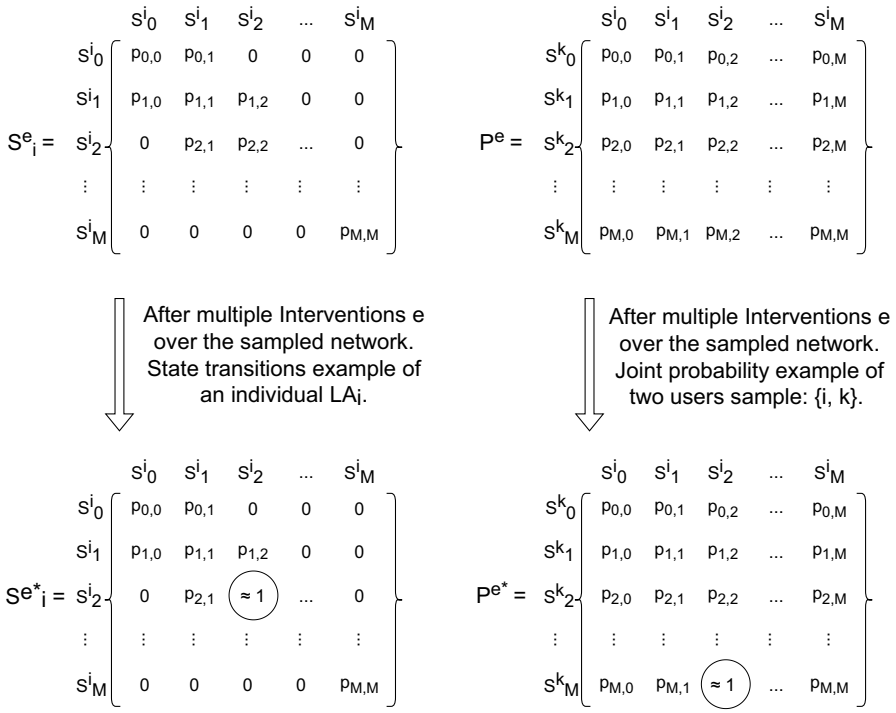

Fig. 10 The individual LA_i state transitions probabilities matrix S_i and the whole LAs joint probabilities matrix P of their joint state transitions from the intervention step e until convergence in intervention step e^*

from the latter three temporal patterns, we establish more realistic information diffusion and mitigation models.

Future work should include a self-supervised detection [40] of the different temporal events instead of the manual annotation. Moreover, more verification techniques should be studied to ensure realistic obtained incentives that would help in the real world. Eventually, the information Twitter API could provide to researchers is considered a limitation since the timestamps of likes are not provided, at least until the time this research was conducted.

Appendix A: Control model

The individual LA system associated with each user and the whole network system can be viewed as Markov systems through the state transitions of the LAs and the joint probabilities of the latter. Figure 10 illustrates both the individual LA_i state transitions probabilities matrix S_i and the whole LAs joint probabilities matrix P. Where state transitions and joint probabilities change between an intervention step e until convergence to a steady state and joint probability of being in

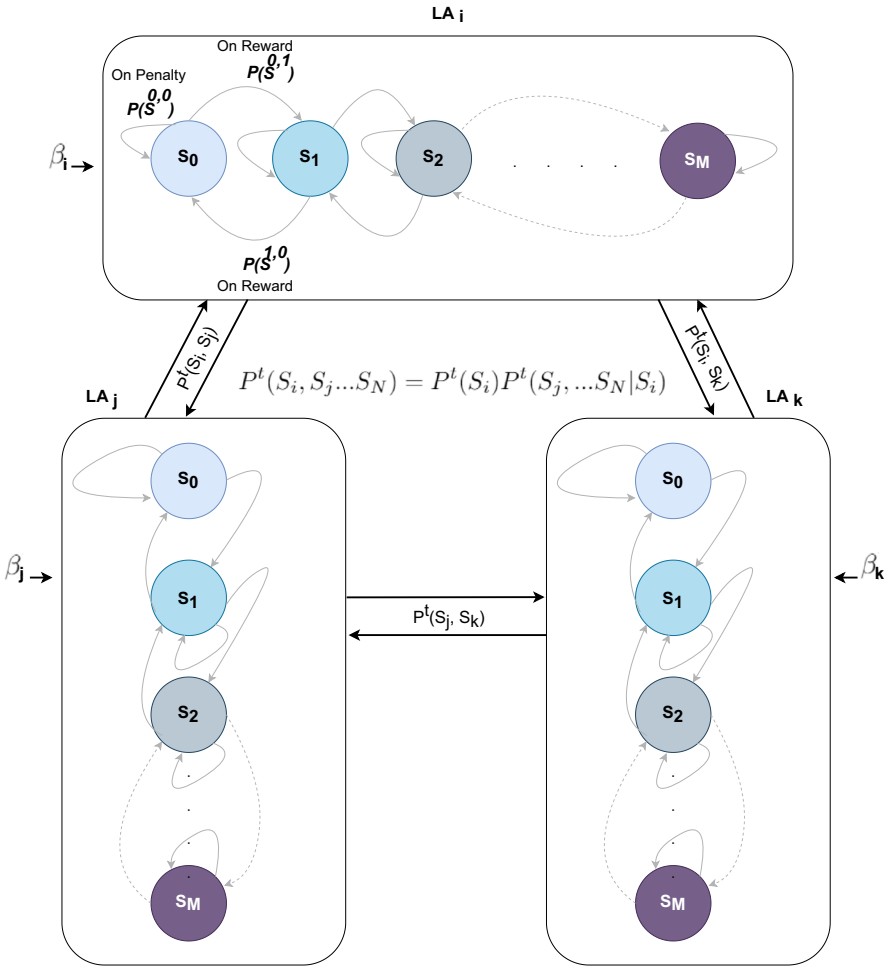

Fig. 11 A toy example of the sampled network of three users and their associated three automata

such states by the intervention step e^*. As indicated in Fig. 10, each LA_i can only perform a state transition by moving one step either to the left, right, or staying at the current state. Eventually, the state transition convergence for each LA_i means it converged to a transition probability of staying at its current state with a value close to 1.

Further, the joint probability of being in a particular state for all LAs determines the final incentive values on a sampled network. These state transitions are governed by a reward and penalty signal β as shown in Fig. 11. Such a signal comes from evaluating the gradient of the total loss function. For instance, if the total loss declined compared to its previous value, the LA that caused that will be rewarded, and its state transition will be committed. Otherwise, it will be

penalized and should stay in its current state. We adopted the same probability calculations and reward function of the utilized LAs as proposed in [9].

Appendix B: MHP simulations setup

Bias-towards

In such a Multivariate Hawkes Process (MHP), we simulated the temporal bias activity of all the 940 users. That means we trained an MHP with timestamps and event counts aggregated over the time realizations period of the samples labeled as *bias=1*. Hence, for each user and time realization (e.g., 180 minutes), all timestamps within the time realization window were structured accordingly. For instance, the 1^{st} time realization had only the timestamps for events where their Twitter creation times were within the first three hours of the 24^{th} of March 2018. Accordingly, the 2^{nd} time realization contained the bias-towards samples timestamps that occurred between 3:00 AM to 5:59 AM on the same day. Then, we kept shifting the time realizations and their associated timestamps the same way until day 27^{th} 8:59 PM, where the following three hours of the day were not part of the MHP training since they were left for testing the exact three-hours predictions.

We set the decay factor for this MHP to 0.6. The primary purpose of such MHP was to predict all users' bias-towards activity to calculate the initial probability of being biased toward the election. The probabilities for each user were calculated according to the frequency of having an associated event belonging to the label *bias=1* in training and predicted data. This process was not incentivized and was only created to calculate such probabilities for the optimization loss function domain (see Eq. 4, Eq. 9, and Eq. 10).

Bias-against

The same concept of the Bias-towards-MHP training also applies to this process. Therefore, we established it to predict all the 940 users' activity for the bias-against event category to calculate the initial probabilities of users being biased against the election (see Eq. 4, Eq. 9, and 10).

We set the decay factor to 0.7. It is essential to highlight that this process was not incentivized. Alternatively, we established the same MHP event category as discussed below but only for a sampled network (100 users), where that process was incentivized.

Bias-against-sampled

This process had the same training timestamps concept as the above two MHPs. The difference between a sampled bias-against-MHP and the 940 bias-against-MHP is

that the latter was used to estimate the initial probabilities of being against the election. At the same time, the former was essential to predict these probabilities after intervention and assigning the incentives. That means some of these probabilities would change, indicating how good the incentives were for some users for mitigating the bias towards the manipulating election campaign and optimizing the loss function (see Eq. 4, Eq. 9, and Eq. 10). Therefore, this process was incentivized with the amounts of state values from the converged LAs. It had only 100 users since it was part of the simulated MHPs during the Monte Carlo simulation. Hence, we intervened with the sampled network and evaluated each user inside it for the associated LA state value because our proposed non-propaganda incentivization also presented a bias-against concept.

We set the decay factor to 0.9 in this process. Since this is a sampled network of users, this process was repeated with different samples for both training and prediction.

Propaganda-sampled

In this process, we have followed the same structure for training the timestamps but only for the sampled 100 users. The process was repeated with a different sample for training and prediction each time. This MHP was not incentivized since we did not wish to intensify the propagation of political propaganda and was only simulated to obtain the predicted counts for users. The obtained counts were used in the ratio parameter for the optimization loss function (see Eq. 4 and Eq. 6). We set the decay factor for this process to 0.9.

Non-propaganda-sampled

Similar to the propaganda-sampled MHP, we established a repeated MHP for the non-propaganda event category where a random sample represented the timestamps for training and predicting the activity of the sampled users (100 users). This process was incentivized with the incentive amounts from the current evaluated user's associated LA state value.

This process had a decay factor of 0.6. The predictions in this process were the direct outcome of the intervention procedure and assigning of an incentive for the current examined user during the Monte Carlo simulation. Therefore, the event counts were evaluated as part of the ratio parameter in the optimization loss function (see Eq. 4 and Eq. 6).

Societal Circle A

To predict all users' activity on the network on how they engaged with the societal circle A, we established this MHP on all the 940 users to predict their future

generated events for that circle. The same training timestamps structure was adopted for the MHP with a decay factor of 0.9. Since this is not a sampled network MHP, we ran it only once to be able to predict the initial probability of engagement with circle A (see Eqs. 4 and 9) for each user without any incentivization.

Societal Circle B

The exact purpose of the societal circle-A-MHP was adopted in this process since all societal circles on all the 940 users must be predicted to calculate the initial probabilities of being in a specific societal circle. Since this process was used to calculate the initial probabilities, we did not apply any incentivization. We assigned the decay factor for this process with 0.9.

Societal Circle B-sampled

The only main difference between the sampled societal circle-B-MHP and the non-sampled circle-B-MHP is the repetition and incentivization in the former. This MHP was incentivized since we wanted to break other circles by intensifying the engagement with it in addition to agreeing. If the latter had occurred, we would increase the probability of engaging with circle B and agreeing with what it represents, which optimizes our loss function (see Eqs. 4 and 9). This process followed the exact configurations in the non-sampled societal circle-B-MHP.

Societal Circles C, E, F

These three MHPs are identical to the non-sampled circle-A and non-sampled circle-B MHPs. Since we had to consider, all circles predicted counts to calculate the initial probabilities of engaging with a circle. The only difference was the decay factors used as we assigned them with 0.75, 0.9, and 0.9, respectively.

Eventually, we ignored simulating circle D since it had only two users, which was not enough to train a MHP but did not impact the results.

Appendix C: Simulation results

Figure 12 gives an example of some simulations with 100 users' real versus predicted event counts.

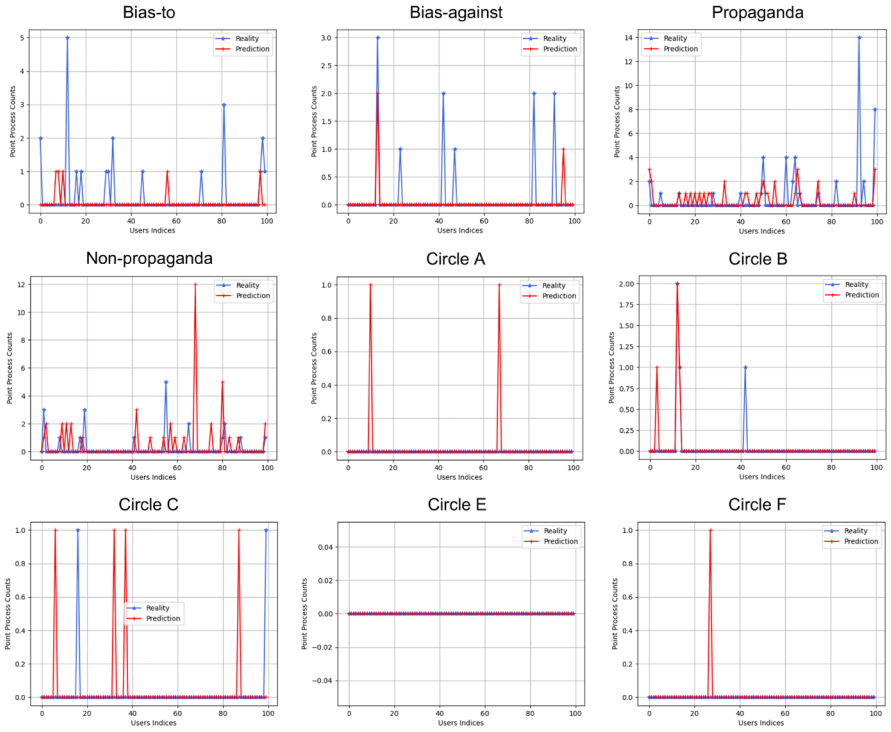

Fig. 12 An example of some simulations with 100 users' real versus predicted event counts

Funding Open access funding provided by University of Agder.

Data Availibility Statement The anonymised data collected are available as open data via PEGYPT [41] GitHub data repository(https://github.com/Ahmed-Abouzeid/PEGYPT).

Declarations

Conflict of interest: On behalf of all authors, the corresponding author states that there is no Conflict of interest.

References

1. Olan, F., Jayawickrama, U., Arakpogun, E.O., Suklan, J., & Liu, S. (2022). Fake news on social media: the impact on society. Information Systems Frontiers, 1–16
2. Farajtabar, M., Du, N., Gomez Rodriguez, M., Valera, I., Zha, H., & Song, L. (2014). Shaping social activity by incentivizing users. Advances in neural information processing systems, *27*
3. Farajtabar, M., Ye, X., Harati, S., Song, L., & Zha, H. (2016). Multistage campaigning in social networks. Advances in Neural Information Processing Systems, *29*
4. Shu, K., Wang, S., Lee, D., & Liu, H. (2020). Mining disinformation and fake news: Concepts, methods, and recent advancements. In: Disinformation, Misinformation, and Fake News in Social Media (pp. 1–19). Springer
5. Abouzeid, A., Granmo, O.-C., Webersik, C., & Goodwin, M. (2021). Learning automata-based misinformation mitigation via hawkes processes. *Information Systems Frontiers, 23*(5), 1169–1188.
6. Farajtabar, M., Yang, J., Ye, X., Xu, H., Trivedi, R., Khalil, E., Li, S., Song, L., & Zha, H. (2017). Fake news mitigation via point process based intervention. In: International Conference on Machine Learning (pp. 1097–1106). PMLR
7. Xu, X., Deng, K., & Zhang, X. (2022). Identifying cost-effective debunkers for multi-stage fake news mitigation campaigns. In: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (pp. 1206–1214)
8. Hair, E., Pitzer, L., Bennett, M., Halenar, M., Rath, J., Cantrell, J., Dorrler, N., Asche, E., & Vallone, D. (2017). Harnessing youth and young adult culture: Improving the reach and engagement of the truth® campaign. *Journal of Health Communication, 22*(7), 568–575.
9. Abouzeid, A., Granmo, O.-C., Webersik, C., & Goodwin, M. (2022). Socially fair mitigation of misinformation on social networks via constraint stochastic optimization. arXiv preprint arXiv:2203.12537
10. Li, M., Wang, X., Gao, K., & Zhang, S. (2017). A survey on information diffusion in online social networks: Models and methods. *Information, 8*(4), 118.
11. Granmo, O.-C., & Oommen, B. J. (2010). Optimal sampling for estimation with constrained resources using a learning automaton-based solution for the nonlinear fractional knapsack problem. *Applied Intelligence, 33*(1), 3–20.
12. Goindani, M., & Neville, J. (2020). Social reinforcement learning to combat fake news spread. In: Uncertainty in Artificial Intelligence (pp. 1006–1016) . PMLR
13. Schuster, T., Schuster, R., Shah, D.J., & Barzilay, R. (2019). Are we safe yet? the limitations of distributional features for fake news detection. arXiv preprint arXiv:1908.09805
14. Liu, X., Nourbakhsh, A., Li, Q., Fang, R., & Shah, S. (2015). Real-time rumor debunking on twitter. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (pp. 1867–1870)
15. Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B.J., Wong, K.-F., & Cha, M. (2016). Detecting rumors from microblogs with recurrent neural networks
16. Wang, W.Y. (2017). " liar, liar pants on fire": A new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648
17. Salem, F. K. A., Al Feel, R., Elbassuoni, S., Jaber, M., & Farah, M. (2019). Fa-kes: A fake news dataset around the syrian war. *Proceedings of the International AAAI Conference on Web and Social Media, 13*, 573–582.
18. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2018). Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media. arXiv preprint arXiv:1809.01286
19. Garg, S., & Sharma, D.K. (2020). New politifact: a dataset for counterfeit news. In: 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART) (pp. 17–22) . IEEE
20. Eady, G., Nagler, J., Guess, A., Zilinsky, J., & Tucker, J. A. (2019). How many people live in political bubbles on social media? evidence from linked survey and twitter data. *SAGE Open, 9*(1), 2158244019832705.
21. Suny, P., Li, J., Mao, Y., Zhang, R., & Wang, L. (2018). Inferring multiplex diffusion network via multivariate marked hawkes process. arXiv preprint arXiv:1809.07688
22. Bruns, A. (2019). Filter bubble. Internet Policy Review *8*(4)

23. Kitchens, B., Johnson, S.L., & Gray, P. (2020). Understanding echo chambers and filter bubbles: The impact of social media on diversification and partisan shifts in news consumption. MIS Quarterly *44*(4)
24. Roozenbeek, J., Van Der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological inoculation improves resilience against misinformation on social media. *Science Advances, 8*(34), 6254.
25. Moravec, P., Minas, R., & Dennis, A.R. (2018). Fake news on social media: People believe what they want to believe when it makes no sense at all. Kelley School of Business research paper (18-87)
26. Al-Asadi, M.A., & Tasdemir, S. (2022). Using artificial intelligence against the phenomenon of fake news: a systematic literature review. Combating Fake News with Computational Intelligence Techniques, 39–54
27. Hangloo, S., & Arora, B. (2022). Content-based fake news detection using deep learning techniques: Analysis, challenges and possible solutions. In: 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT) (pp. 411–417). IEEE
28. Khattar, D., Goud, J.S., Gupta, M., & Varma, V. (2019). Mvae: Multimodal variational autoencoder for fake news detection. In: The World Wide Web Conference (pp. 2915–2921)
29. Amer, E., Kwak, K.-S., & El-Sappagh, S. (2022). Context-based fake news detection model relying on deep learning models. *Electronics, 11*(8), 1255.
30. Shu, K., Wang, S., & Liu, H.(2019). Beyond news contents: The role of social context for fake news detection. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (pp. 312–320)
31. Goldman, A. I., & Baker, D. (2019). Free speech, fake news, and democracy. *First Amendment Law Review, 18*, 66.
32. Wang, S., Xu, X., Zhang, X., Wang, Y., & Song, W. (2022). Veracity-aware and event-driven personalized news recommendation for fake news mitigation. In: Proceedings of the ACM Web Conference 2022 (pp. 3673–3684)
33. Kobayashi, R., & Lambiotte, R. (2016). Tideh: Time-dependent hawkes process for predicting retweet dynamics. In: Tenth International AAAI Conference on Web and Social Media
34. Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., et al. (2020). Bias in data-driven artificial intelligence systems-an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(3), 1356.
35. Amburgey, T. L. (1986). Multivariate point process models in social research. *Social Science Research, 15*(2), 190–207.
36. Rizoiu, M.-A., Lee, Y., Mishra, S., & Xie, L. (2017). A tutorial on hawkes processes for events in social media. arXiv preprint arXiv:1708.06401
37. Yazidi, A., Bouhmala, N., & Goodwin, M. (2020). A team of pursuit learning automata for solving deterministic optimization problems. *Applied Intelligence, 50*(9), 2916–2931.
38. Raychaudhuri, S. (2008). Introduction to monte carlo simulation. In: 2008 Winter Simulation Conference (pp. 91–100). IEEE
39. Bankes, S., Lempert, R., & Popper, S. (2002). Making computational social science effective: Epistemology, methodology, and technology. *Social Science Computer Review, 20*(4), 377–388.
40. Abouzeid, A., Granmo, O.-C., Goodwin, M., & Webersik, C. (2022). Label-critic tsetlin machine: A novel self-supervised learning scheme for interpretable clustering. In: 2022 International Symposium on the Tsetlin Machine (ISTM) (pp. 41–48). IEEE
41. Abouzeid, A. Ahmed-Abouzeid/PEGYPT: V1.0.0. https://doi.org/10.5281/zenodo.7780594 .

Authors and Affiliations

Ahmed Abouzeid[1] · Ole-Christoffer Granmo[1] · Morten Goodwin[1] · Christian Webersik[2]

✉ Ahmed Abouzeid
ahmed.abouzeid@uia.no
https://scholar.google.com/citations?user=wpvNoXIAAAAJ&hl=en

Ole-Christoffer Granmo
ole.granmo@uia.no
https://scholar.google.com/citations?hl=en&user=PmdKAykAAAAJ

Morten Goodwin
morten.goodwin@uia.no
https://scholar.google.com/citations?hl=e&user=da-byPgAAAAJ

Christian Webersik
christian.webersik@uia.no
https://scholar.google.com/citations?hl=en&user=wo2fq9sAAAAJ

[1] Center for Artificial Intelligence Research, University of Agder, Jon Lilletuns vei, Grimstad 4879, Agder, Norway

[2] Center for Integrated Emergency Management, University of Agder, Universitetsveien, Kristiansand 4630, Agder, Norway