EFFECTS OF HIDDEN OPINION MANIPULATION IN MICROBLOGGING PLATFORMS

GIULIA BRAGHINI AND FRANCESCO SALVARANI

Abstract. This article studies a model for describing opinions' evolution in a community linked by microblogging directed network system. The network can evolve in time, by means of the creation and the deletion of some connections. When a connexion is created, the individuals have access to the whole history of posts written by the corresponding author and, when a connexion is destroyed, all the posts written by the corresponding author become invisible. The agents' opinions are described by a set of continuous opinion variables in the closed interval $\Omega = [-1, 1]$. They represent the agreement (or disagreement) of the corresponding agent with respect to a binary question (such as a referendum or an election with two candidates). The model takes into account the effects on public opinion caused by the sign and the intensity of the initial opinions of the agents, their activity in microblogging platforms the presence of leaders and the possible manipulations of the visibility of the posts by the microblogging platform owner. The model is written as a system of integro-differential equations and simulated by using a Runge-Kutta method. We show that hidden manipulation can have an important impact on the public opinion formation and that very mild interventions of the network owner may induce major effects on the population. In our simulations, the effect of hidden manipulation is shown to be more efficient in driving the public opinion than the action of a leader, in the case of bounded confidence models and for short-time intervals.

1. Introduction

In the last years, the use of internet, social media and social networks has become very popular because of its immediacy and ease of use and, because of the recent COVID-19 pandemics and the social distancing due to the prevention health policies, the online dissemination of verified or unverified contents has highly increased [18, 28, 30].

The online sharing of information, advices and suggestions may have a major effect on the opinion formation phenomenon. The authors of web-based contents may have several reasons which explain their actions. Many users simply aim to inform the audience about some facts and the author's interpretation and opinion about them, other simply forward news and opinions which seem important to the sender and some others aim to persuade the audience and drive the public opinion towards a given direction [22]. In many cases, the authors are known, but sometimes they are not [3].

By looking at the contents written with persuasion purposes, several strategies can be adopted. A strategy, which is the easiest to be detected, consists in the explicit declaration by a known user of his\her viewpoint. The content develops all the arguments which support the thesis of its author, whose identity is known. However, the public opinion can also be driven by means of other techniques and several individuals may be the object of hidden manipulation, with possible non-negligible consequences on the collective behaviour [27].

A very well-known case is the so-called *Facebook-Cambridge Analytica* data scandal [26], which highlighted some possible consequences of social media sharing, especially in connexion with non-apparent opinion manipulation techniques related to elections or referendums.

For these reasons, much attention has recently been paid to the power of social media and social networks in the dynamics of collective choices, especially in the political and commercial frameworks.

Date: October 8, 2021.

The interest for these questions fits into a well established line of research, already active for many years, which has led many authors to propose specific mathematical models for describing the emergence of consensus in interacting groups. The main idea underlying this line of research is the study of social influence as a micro-level process for deducing macro-consequences for consensus or divisions in society.

A possible systematization of the contributions published up to now is based on the mathematical properties of the models (see [29] and the references therein). In particular, the main ingredient of all models is the opinion space, which can be discrete, continuous over a bounded interval or continuous over $\mathbb{R}$. The interaction can be pairwise, any-to-any or with respect to the closest neighbours. The time variable can be discrete or continuous. The model can be written under the form of a system of difference equations (see, for example, [16, 17, 23]), a system of ordinary differential equations [1, 15, 12], a kinetic equation (e.g. [4, 7, 8, 9, 11, 35]), a partial-differential equation [10, 35]. The Markov chain viewpoint has also been proposed [6].

In [20], the authors provide an extensive analysis on the main modelling hypotheses and give an extensive literature on the subject. In particular, they classify the reviewed models into three classes: models of assimilative social influence, models with similarity biased influence and models with repulsive influence.

Often, the outputs of the models confirm, on the one hand, many expected dynamics. On the other hand, some models also predict certain counterintuitive behaviours. An example is given by a recent general bounded confidence model proposed by Hegselmann and Krause in [24] for a population influenced by their normal peers, exposed to an additional external source of influence, such as radical groups or charismatic leaders. Their model foresees, for example, that stronger signals may have less effect than weaker ones or that more radicals may reduce the radicalization of the members of a population. Other studies have shown that a richer dynamics can be deduced by taking into account external influence sources or heterogeneous populations. For instance, [6] showed how restrictions in communication leading to the co-existence of different opinions follow from the emergence of new absorbing states and [25] studies how the interaction between open-minded and closed-minded groups induces a novel class of equilibria consisting of multiple connected opinion clusters.

An individual opinion has social effects only if it is disseminated within the community itself. For this reason, the medium of transmission is important. In this paper, we will specifically study the effects of microblogging and social networking services, on which users may post messages and interact with other users, as they have proven to be an efficient tool for conveying opinions and they may have an effect in influencing the public opinion [19].

In this article, we aim to understand the effects of two phenomena which induce opinion modifications, which we call hidden and explicit influencing. For the purposes of the article, explicit influence is the effect on the population's opinion by a known influencer, whereas hidden influence is the effect of the platform owner in highlighting or in hiding information.

The sociological literature agrees that one of the main dynamics of opinion formation is *consensus* [32, 34] and, consequently, many mathematical models of social interaction, such as the well-known Hegselmann-Krause model [23], the Deffuant-Neau-Amblard-Weisbuch model [17] and the Cucker-Smale model [16] are based on consensus dynamics if the confidence bound is large enough.

Opinion formation being mainly driven by consensus, it is clear that the neutrality of the blogging platform a crucial factor. Indeed, if the platform is not neutral (for example, by hiding some posted opinions or by ordering them in a deliberated way), then the process of opinion formation is biased.

One of the goals of this article is the study of a particular technique of manipulation, which does not hide any posts, but rather ranks them in such a way that only some tendencies are highlighted (for example, by means of an ordering algorithm based on AI semantic algorithms). Note that this technique is more sophisticated than the usual strategy based on opinion manipulation bots, i.e. software applications that automatically produce posts on the

web [36], because it works with trusted posts of real individuals, often directly known by the reader. This technique requires, however, that the owner of the microblogging platform network is involved in the manipulation dynamics. This phenomenon is known as *filter bubble:* it results in the partial visibility of a global situation caused by the selection of online contents by an algorithm [31]. In fact, search engines often use known information about the user (such as location, search history, online behaviour) to select and classify information to show [13]. Information filtering processes take place on the individual, the social, and the technological levels: this behaviour is known as the triple-filter-bubble framework [21].

We provide a model which shows in a quantitative way the evolution of the opinions in a closed community by taking into account two main features of the agents, their post productivity and their opinions, as well as the evolution of the network between the individuals, the presence of an external influencer and the policy of the blogging platform in visualizing the posts among the interested users.

The model presented in this article cannot be simply attributed to one of these classes indicated in [20], even if assimilative social influence plays a major role. Our model, which generalizes the model described in [24], has the following features.

- It takes into account in a unified way the effect of an external agent, which can be either the platform owner or an external influencer, whose action can considerably modify the evolution of the final state of the system, the time evolution of the network and the productivity of the agents. In this way, both hidden manipulation and public persuasion actions can be described by the same model and their effects on the opinion evolution can be compared. The unified approach is possible through a set of highlighting functions, which can model both the effect of an explicit external agent and the hidden manipulation obtained through a selection criterion which modifies the visibility of the opinions expressed within the network.
- It takes into account the agent's memory. Most models of opinion formation, including the model introduced in [24], are local in time. The introduction of memory terms has an effect on the opinion evolution and implies the use of specific mathematical and numerical strategies.
- The structure of the model is well adapted for studying existence and uniqueness of the solution by using sophisticated mathematical tools.

We have chosen to work in a continuous-time framework. This choice has several advantages, mainly at the theoretical level. It allows, on the one hand, to apply the theory of regular Lagrangian flows (see Definition 3.1 and Theorem 3.2) to individuate under which conditions the system admits a unique solution. On the other hand, it opens the way to the rigorous study of the long-time asymptotics of the model, especially the relaxation speed to equilibrium, by using recent mathematical techniques, such as those described in [2] and in [12]. At the numerical level, we have used a discretization based on Runge-Kutta routines, which is consistent with the continuous model.

In order to highlight the importance of this step, we give a counterexample which shows that the property of uniqueness of the solution may fail even in the case of an apparently reasonable model.

The outputs of the model show that hidden manipulation is a very efficient tool for influencing a population. In particular, we have compared the effects of hidden manipulation and the explicit influence of a leader in order to drive the population's consensus towards a given value. Our simulations show that, under the circumstances of our tests, hidden manipulation can be a more efficient strategy for modifying the final consensus with respect to the action of an external leader, at least on finite time intervals. Our simulations show moreover that the fraction of individuals with positive opinion exhibits high-frequency oscillations, which may be seen as similar to random effects in the evolution of such a quantity. However, as we will detail when describing the numerical simulations, the effect is purely deterministic.

The structure of the article is the following. In Section 2, we describe our mathematical model. Then, in Section 3, we propose our counterexample and discuss the well-posedness of

our model. Section 4 describes and discusses several numerical results, including the stability of the manipulation process with respect to the number of connexions within the network. Because of its intrinsic interest in opinion dynamics and of its practical applications, the discrete-time version of the model is also meaningful. We provide its detailed description in the Appendix.

2. Description of the mathematical model

We consider a population composed of $N \in \mathbb{N}^*$ interacting individuals, described – at the individual level – by N time-dependent functions

$$x_i \ : \ \mathbb{R}^+ \to \Omega = [-1, 1] \qquad i = 1, \dots, N.$$

The functions x_i $(i = 1, \dots, N)$ represent the opinion of the agent labelled with the index i with respect to a binary question (which can be, for example, a referendum, an election with two candidates or the opinion with respect to a commercial product). When $x_i = 1$, the i-th agent completely agrees with the underlying question whereas, when $x_i = -1$, the i-th agent is in full disagreement with the underlying question. All intermediate values belonging to the open interval $(0, 1)$ denote partial agreement with the binary question, with a conviction proportional to the magnitude of x_i and, symmetrically, the intermediate values belonging to the open interval $(-1, 0)$ denote partial disagreement with the binary question, with a conviction proportional to the absolute value of x_i. When $x_i = 0$, the agent has no preference about the binary question.

We suppose that the individual opinion is publicly available under the form of posts in the blogging platform and that it evolves only through reciprocal influence. The individuals of the population are totally or partially interconnected by means of an oriented graph, represented by a set of time-dependent functions of binary type

$$\sigma_{i,j} \ : \ \mathbb{R}^+ \to \{0, 1\} \text{ for all } i, j = 1, \dots, N.$$

If the i-th agent is following the j-th agent at time $t \in \mathbb{R}^+$, then $\sigma_{i,j}(t) = 1$, otherwise $\sigma_{i,j}(t) = 0$. The matrix whose entries are the quantities $\sigma_{i,j}$ will be denoted, in the whole article, as the *interaction matrix*.

In what follows, we suppose that the population is interconnected in such a way that, for all $i = 1, \dots, N$ and for all $t \in \mathbb{R}^+$, there exists at least an index $j \neq i$ such that $\sigma_{i,j} = 1$ (it means that no agent is fully isolated). We suppose moreover that each agent has a total access to his\her own posts: $\sigma_{i,i} = 1$, for all $t \in \mathbb{R}^+$ and for all $i = 1, \dots, N$. It is important to underline that, like in real social media, the interaction matrix is often sparse.

We moreover denote with $b_i = b_i(t)$ the number density, with respect to t, of microblogs posted by the i-th individual.

Our model aims to forecast the opinion evolution on a short-time horizon (for example the dynamics of a referendum campaign). Consequently, we can assume that there is no loss of attention about the underlying question.

The set of ordinary differential equations of our model describes the opinion evolution through a consensus dynamics, and takes into account the activities of the agents as microbloggers. Its precise form is the following:

$$\begin{cases} \dfrac{db_i}{dt}(t) = \mu_i(\gamma_i - b_i(t)) \displaystyle\sum_{j=1}^{N} \sigma_{i,j}(t) \int_0^t b_j(\theta)\, d\theta, \qquad \gamma_i,\ \mu_i > 0 \\ \\ \dfrac{dx_i}{dt}(t) = \alpha_i(x_i(t))(\Phi_i(t) - x_i(t)), \end{cases} \tag{2.1}$$

where

$$\Phi_i(t) = \begin{cases} \dfrac{\sum_{j=1}^{N} \sigma_{i,j}(t) \int_0^t b_j(\theta) x_j(\theta) \psi_{i,j}(\theta) \mathrm{d}\theta}{\sum_{j=1}^{N} \sigma_{i,j}(t) \int_0^t b_j(\theta) \psi_{i,j}(\theta) \mathrm{d}\theta} & t > 0 \\ \dfrac{\sum_{j=1}^{N} \sigma_{i,j}(0) b_j(0) x_j(0) \psi_{i,j}(0)}{\sum_{j=1}^{N} \sigma_{i,j}(0) b_j(0) \psi_{i,j}(0)} & t = 0 \end{cases} \tag{2.2}$$

and

$$\psi_{i,j} = \begin{cases} \psi_{i,j}(\theta, x_j(\theta), b_j(\theta)) > 0 & i \neq j \\ 1 & i = j. \end{cases} \tag{2.3}$$

The model is coupled with suitable initial conditions: for all $i = 1, \dots, N$

$$b_i(0) = b_i^0 \in (0, \gamma_i), \tag{2.4}$$

$$x_i(0) = x_i^0 \in \Omega. \tag{2.5}$$

We now describe each term of the model.

The equations satisfied by the functions $b_i(t)$ are of logistic type. We indeed suppose that the activity of the i-th microblogger is proportional to the number of total microblogs seen by him\her up to a saturation phenomenon The parameters γ_i and μ_i represent the posts' production saturation values and the logistic growth rates for the i-th individual. We are indeed considering a short-time model, which means that we do not expect any interest decrease about the underlying question.

In our model, when $\sigma_{i,j} = 0$, the i-th agent loses all the posts sent by the j-th agent, but he\she will see again all the posts of the j-th agent as soon as $\sigma_{i,j} = 1$, as customary in many microblogging platforms.

The equations describing the time behaviour of the opinions $x_i(t)$, $i = 1, \dots, N$, are of consensus type. We suppose that the i-th agent modifies his\her own opinion through a consensus dynamics by taking into account the opinions of the individuals he\she is following. In (2.1), the time evolution of the functions x_i is governed by the joint contribution of two terms. The functions $\alpha_i : \Omega \to \mathbb{R}^+$ are somehow the analogous of the admissible functions defined in [9] (Definition 2.6): they may be agent-dependent and translate the idea that individuals with a stronger opinion are more stable in their convictions. In general, we suppose that all the α_i are even functions (because of the symmetry under the exchange of the underlying question with its opposite) and of class $W^{1,1}(\Omega)$.

The variation of the opinion for the i-th agent with respect to time, at time t, is given by the difference $(\Phi_i - x_i)$, weighted by the term α_i.

The functions Φ_i describe a weighted average opinion of the posts seen by the i-th agent. We suppose that all the posts of the followed individuals are available and that a post of an agent at a given time strictly reflects his\her opinion at the same time.

We suppose that the weighted opinion deduced by the set of posts available to the i-th agent is given by the integral in time of all the posts sent by all the individuals followed by the i-th agent, weighted with suitable *highlighting functions* $\psi_{i,j}$.

The highlighting functions, in a non-manipulated environment, may describe the memory of the agents. If the posts are ordered only with respect to time, a plausible form of these highlighting functions may be a negative exponential in time or the characteristic function of a suitable subset of the interval $(0, t)$. This assumption reflects the idea that more recent posts have a greater influence on the readers than older posts. However, in a manipulated situation, the highlighting functions describe the possible manipulation induced by the owner

of the platform. Since the highlighted posts are chosen by the owner of the platform in order to maximise their effect, it is therefore important to allow the platform owner to consider the whole history of posts, including the oldest ones.

Note that timing is a crucial factor in opinion formation dynamics (see [7] for a discussion about this point): for this reason we allow that the manipulation strategy may vary in time.

In both cases, these quantities are normalized by the total number of posts, weighted by the highlighting functions. Note that the quantities $\sigma_{i,j}$ in front of this weighted average (which guarantee that only the agents followed by the i-th individual are taken into account in this average) are considered at time t. It means that, when a user follows another agent, he\she has a complete access to all his\her posts, and, when a user decide to eliminate another individual from his\her set of contacts, he\she looses the access to all his\her comments.

As said before, the highlighting functions $\psi_{i,j}$, in a manipulated situation, describe the manipulation effect. These terms are supposed to be under the control of the microblogging network's provider. When there is no manipulation, the $\psi_{i,j}$ depend only on t for all $i, j = 1, \dots, N$. In the case of hidden manipulation, we suppose that $\psi_{i,i}$ depend only on the time variable for all $i = 1, \dots, N$ for preventing the individual to see any manipulation effect on his\her own posts. Moreover, we suppose that $\psi_{i,j}(t) \in (0, 1]$ for all $t \in \mathbb{R}^+$ (hence, $\psi_{i,j} \in L^\infty(\mathbb{R}^+)$ for all $i, j = 1, \dots, N$). In a manipulated situation, the highlighting functions may depend on time, on the opinion of the agents followed by the i-th agent and of the number of posts seen by him\her.

We do not allow complete censorship; consequently, we suppose $\psi_{i,j} > 0$ in (2.3). A possible implementation of a manipulation technique consists in simply promoting, in the ranking of posts, those which are favourable to the thesis supported by the manipulator and by postponing in the ranking the unfavourable posts. The effectiveness of this technique is the consequence of the information overload of microblogging platforms [33]: usually not every post is read, especially when their number is high, and the reader limit himself to the first ones. As we will see in the next sections, the effect of the highlighting functions may result in a modification of the asymptotic state of the system.

If we suppose that $\psi_{i,j} = 1$ for all $i, j = 1, \dots, N$, and all the α_i are constant, we obtain a linear system of ODEs for the unknowns x_i of Hegselmann-Krause type [23].

Without manipulation phenomena, the evolution of the population is the consequence of several factors inside the population. The number of individuals with opinion of the same sign is, of course, important and is the goal of the majority of polls. However, it is not enough for explaining the evolution of the population's global opinion, because at least two other factors are of paramount importance: the activity in sharing their opinions – here measured by means of the individual number of posts b_i – and the conviction degree of each agent, which corresponds to the absolute value of his/her opinion, $|x_i(t)|$. The implementation of polls with multiple answers about a binary question, modulated on a scale, is hence very useful for producing accurate forecasts (see, for example, [11]).

Leaders are modelled by individuals with peculiar forms of the highlighting functions and of the corresponding entries of the interaction matrix. By supposing the existence of only one leader in the population (for simplicity labelled by the index N), the entries of the interaction matrix involving the leader have the form

$$\sigma_{N,i} = 0, \quad \sigma_{i,N} = 1 \text{ for all } i = 1, \dots, N-1$$

and the highlighting functions are such that $\psi_{i,N} \gg \psi_{i,j}$ for all $i = 1, \dots, N-1$ and $j = 1, \dots, N-1$.

3. Basic mathematical properties of the model

This section is devoted to the mathematical analysis of our model. In particular, we will discuss the well-posedness of the model (i.e. the existence and the uniqueness of the solution), which is key feature of any good mathematical model.

In particular, we underline that the lack of uniqueness is a serious limitation to the predictive value of any mathematical model. A considerable advantage of the choice to model the phenomenon with a system of multi-agent differential equations is precisely the possibility of studying, with sophisticated techniques of mathematical analysis, the conditions that guarantee the uniqueness of the solution of the model.

3.1. **A counterexemple.** We first describe a counterexample that emphasizes the importance of well-posedness in opinion formation multi-agent systems.

Consider a population composed by N interacting agents. We introduce a $N \times N$ matrix S whose entries, denoted with $\sigma_{i,j}$, describe the effects of the influence of the agent labelled with the index i on the agent labelled with index j. In what follows, we suppose that $\sigma_{i,j}$ may assume only two values, zero and one. When $\sigma_{i,j} = 0$, we suppose that there exist no interaction between the corresponding individuals and, when $\sigma_{i,j} = 1$, the i-th individual has an influence on the opinion of the j-th individual.

Let $x_i = x_i(t) \in$ the opinion of the i-th agent. For simplicity, we will work in an unbounded domain, hence $x_i \in \mathbb{R}$ for all $i = 1, \dots, N$.

The equations of the counterexample are the following. Let

$$\phi(y) = \begin{cases} \sqrt{y} - y & y \geq 0 \\ -\sqrt{|y|} - y & y < 0. \end{cases} \tag{3.1}$$

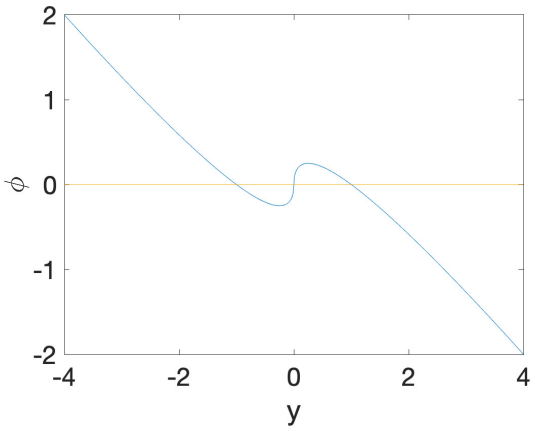

FIGURE 1. Plot of the graph of the function $y \mapsto \phi(y)$.

The evolution of x_i with respect to time is given by the set of differential equations

$$x_i'(t) = \frac{1}{2} \sum_{i=1}^{N} \sigma_{i,j} \phi(x_i - x_j), \qquad i = 1, \dots, N. \tag{3.2}$$

This set of equations models a system which tends to consensus when the starting opinions of the agents are such that $|x_i - x_j| > 1$. When individuals interact with other agents having opinions very similar to their own, the binary encounters induce different shades of opinion.

The equations hence describe a model of optimal distinctiveness, which reconciles the opposing needs of assimilation and differentiation from others. This behaviour has been analyzed in the literature [14].

We now suppose that

- $N \geq 2$ is an even natural number;
- $\sigma_{2i-1,2i} = 1$ and $\sigma_{2i,2i-1} = 1$ for all $i = 1, \dots, N/2$ and zero otherwise;
- all the initial opinions of the agents are equal to zero, i.e.

$$x_i'(0) = 0, \qquad i = 1, \dots, N. \tag{3.3}$$

System (3.2) can hence be decoupled in $N/2$ independent 2×2 systems having the form

$$\begin{cases} x'_{2i-1}(t) &= \dfrac{1}{2}\phi(x_{2i-1} - x_{2i}) \\ x'_{2i}(t) &= \dfrac{1}{2}\phi(x_{2i} - x_{2i-i}) \end{cases}$$

with initial conditions $x'_{2i-1}(t) = 0$ and $x'_{2i}(t) = 0$ for all $i = 1, \dots, N/2$. Thanks to the parity property of ϕ, if we subtract the second equation from the first one, we end with

$$[x_{2i-1} - x_{2i}(t)]'(t) = \phi(x_{2i-1} - x_{2i}), \qquad i = 1, \dots, N/2. \tag{3.4}$$

The solutions of (3.4) are hence the solutions of the Cauchy problem for the ordinary differential equation

$$z'(t) = \begin{cases} \sqrt{z} - z & z \geq 0 \\ -\sqrt{|z|} - z & z < 0. \end{cases} \qquad z(0) = 0, \tag{3.5}$$

where $z(t) = x_{2i-1}(t) - x_{2i}(t)$ for a given i. It is easy to verify, by direct inspection, that both

$$z(t) = 0 \text{ and } z(t) = (e^{-t/2} - 1)^2$$

are solutions of (3.5) for $t \geq 0$. Hence, system (3.1)-(3.2)-(3.3) does not have a unique solution.

3.2. **Well-posedness analysis.** A crucial step is hence the identification of the conditions on a differential multi-agent system which guarantee the existence and the uniqueness of the solution. This study is detailed in the present subsection.

We first remark that our model is composed of two sets of weekly coupled unknowns: the functions representing the number of posts written by the i-th individual (denoted b_i) and the the functions representing the opinions x_i.

It is possible to decouple the equations satisfied by the unknowns b_i in Equation (2.1): the first family of equations in Equation (2.1), i.e.,

$$\begin{cases} \dfrac{db_i}{dt}(t) = F_i(t, b_1, \dots, b_N(t)) := \mu_i(\gamma_i - b_i(t)) \displaystyle\sum_{j=1}^{N} \sigma_{i,j}(t) \int_0^t b_j(\theta)\, d\theta \\ \\ b_i(0) = b_i^0 \in (0, \gamma) \end{cases} \tag{3.6}$$

can be solved as a coupled system for all the b_i, by applying the standard Cauchy-Lipschitz theory, all the F_i being of class $C^\infty((\mathbb{R}^+)^{N+1})$. We can hence deduce that there exists one and only one vector $b = (b_1, \dots, b_N) \in C^\infty((\mathbb{R}^+)^N)$, solution of system (3.6) with initial data $b_i(0) = b_i^0$.

Consider the second half of system (2.1) with known vector $b \in C^\infty((\mathbb{R}^+)^N)$:

$$\begin{cases} \dfrac{dx_i}{dt}(t) = \alpha_i(x_i(t))(\Phi_i(t) - x_i(t)) \qquad i = 1, \dots, N \\ \\ x_i(0) = x_i^0 \in \Omega. \end{cases} \tag{3.7}$$

All the Φ_i satisfy the following bound for all $T > 0$:

$$0 \leq \Phi_i \leq \max_{j=1,\dots,N} \sup_{t \in [0,T]} |x_j(t)|.$$

As a consequence of the binary character of the interaction matrix and the regularity hypotheses of the highlighting functions $\psi_{i,j} \in L^\infty(\mathbb{R}^+)$, the terms Φ_i – which are linear with respect to the opinion variables – cannot be more regular than L^∞ functions with respect to time. Moreover, the equations are not sufficiently regular with respect to the unknown $x = (x_1, \dots, x_N)$ for applying the Cauchy-Lipschitz theory because of the low regularity of the vector $\alpha = (\alpha_1, \dots \alpha_N)$.

For this reason, we need to base our study on a more general theory (we refer to the lecture notes [5] and to the references therein for a complete introduction to the theory of flows associated to non-smooth vector fields).

The Cauchy problem (3.7) has the following structure:

$$\begin{cases} \dot{x}(t) = \eta(t, x(t)) \\ x(0) = x_0, \end{cases} \tag{3.8}$$

where $x : [0,T] \to \Omega^N$ is the opinion vector for the whole population and $\eta : [0,T] \times \Omega^N \to \mathbb{R}^N$ is the associated vector field, which may have no Lipschitz regularity.

Let $\mathcal{L}^N$ be the N-dimensional Lebesgue measure and consider a map $X : [0,T] \times \mathbb{R}^N \to \mathbb{R}^N$. The fact that $X(t,\cdot)\#\mathcal{L}^N \leq L\mathcal{L}^N$ for all $t \in [0,T]$, where the symbol $\#$ represents the push-forward of a measure, means that there exists $L > 0$ such that, for all $t \in [0,T]$ and for all $\phi \in C_c^0(\mathbb{R}^N) \geq 0$,

$$\int_{\mathbb{R}^N} \phi(X(t,x))\mathrm{d}x \leq L \int_{\mathbb{R}^N} \phi(x)\mathrm{d}x.$$

For a given vector field η, we consider as admissible solutions to the system the maps called regular Lagrangian flows (see [5]):

Definition 3.1. *A regular Lagrangian flow is a map $X : [0,T] \times \mathbb{R}^N \to \mathbb{R}^N$ such that:*

i) for $\mathcal{L}^N$-a.e. $x \in \mathbb{R}^N$, the function $t \to X(t,x)$ is a solution of the ODE in the integral sense, i.e. such that:

$$X(t,x) = x_0 + \int_0^t \eta(s, X(s,x))ds \qquad \text{for all } t \in [0,T];$$

ii) there exists a constant $L > 0$ such that $X(t,\cdot)\#\mathcal{L}^N \leq L\mathcal{L}^N$, for all $t \in [0,T]$.

The constant L is called compressibility constant. The following theorem holds [5]:

Theorem 3.2. *Consider $\eta \in (L^1[0,T]; W^{1,1}(\mathbb{R}^N;\mathbb{R}^N))$, such that $\eta \in L^\infty([0,T] \times \mathbb{R}^N;\mathbb{R}^N)$ and $[div_x \eta]^- \in L^1([0,T]; L^\infty(\mathbb{R}^N))$. Then there exists a unique regular Lagrangian flow X associated to the field η, solution of the Cauchy problem* (3.8).

By using the notation of this article, we immediately deduce:

Theorem 3.3. *Consider the Cauchy problem* (2.1)-(2.5) *for $t \in [0,T]$, $T > 0$. Let $\sigma_{i,j} : \mathbb{R}^+ \to \{0,1\}$ for all $i,j = 1,\dots,N$ be a set of functions of class $L^\infty(0,T)$. Let $\psi_{i,j}(t) \in (0,1]$ for all $t \in \mathbb{R}^+$ for all $i,j = 1,\dots,N$. Suppose moreover that the field $\alpha \in (L^1[0,T]; W^{1,1}(\Omega^N;\mathbb{R}^N))$, $\alpha \in L^\infty([0,T] \times \Omega^N;\mathbb{R}^N)$ and $[div_x \alpha]^- \in L^1([0,T]; L^\infty(\Omega^N))$. Let $b_i^0 \in (0,\gamma_i)$, $\gamma_i > 0$, $\mu_i > 0$ and $x_i^0 \in \Omega$ for all $i = 1,\dots,N$.*

Then, there exists one and only one solution of (2.1)-(2.5). *The opinion vector x is a regular Lagrangian flow and $b \in C^\infty((\mathbb{R}^+)^N)$.*

Moreover, if α is a Lipschitz field with respect to the opinion vector x, uniformly in time, then existence and uniqueness of the solution hold in the classical sense for both b and x.

4. Numerical results

Because of the weak coupling of the model, already described in the previous section, the numerical simulations have been produced by decoupling the problem in two sub-problems.

We introduce the functions

$$B_i = \int_0^t b_i(\theta)\, d\theta, \qquad i = 1,\dots,N, \tag{4.1}$$

which represent the total number of microblogs posted at time t by the individual labelled with the index i. Thanks to the regularity of b, proved in the previous section, we deduce

immediately that (3.6) can be written as a pure differential system:

$$\begin{cases} \dfrac{db_i}{dt}(t) & = \mu_i(\gamma_i - b_i(t)) \displaystyle\sum_{j=1}^{N} \sigma_{i,j}(t) B_j(t) \\ \dfrac{dB_i}{dt}(t) & = b_i(t) \\ b_i(0) & = b_i^0 \in (0, \gamma) \\ B_i(0) & = 0. \end{cases} \tag{4.2}$$

We have first solved the Cauchy problem (4.2) and then we have stocked the results of the problem. We have subsequently used them as input data for solving the Cauchy problem (3.7). Both systems have been discretized by means of a standard fourth-order Runge-Kutta routine.

In what follows, we always suppose that

$$\alpha_i(s) = \beta(1 - s^2), \qquad \beta > 0, \qquad \text{for all } i = 1, \dots, N. \tag{4.3}$$

This specific form of the nonlinear fields α_i is coherent with the hypothesis that individuals with extreme opinions are more stable in their convictions.

We separately treat two network geometries. The first geometry describes a strongly connected network (i.e. there exists a path linking each pair of agents of the population) and the second one describes a partially interconnected network, composed of separate clusters of agents.

The time evolution of the opinions in each geometry is then analyzed by looking at different situations: the first one without hidden manipulation, the second in presence of hidden manipulation. The last geometry presents a comparison between hidden manipulation and explicit influence.

In all the numerical simulations, we consider a population composed of $N = 10^3$ partially interconnected individuals. The choice of this value for N allows to produce readable figures and to work with a population in which an individual behaviour has little effect at the collective level. Of course, simulations with a greater number of agents are possible and do not induce major difficulties, at least when N is not too big.

The time step of the Runge-Kutta algorithm is $\Delta t = 5 \times 10^{-3}$ and the simulations have been displayed for $t \in [0, 5]$, t being measured in weeks. We moreover choose the following numerical values. For all i, $\mu_i = \mu^* = 10^{-4}$: we hence suppose that the post's production is saturated after 3 weeks; the maximum number of daily posts for each agent is set to $\gamma_i = \gamma^* = 10$ and the relaxation constant (which represents the interaction frequency between the agents) in (4.3) is $\beta = 2$.

For each numerical experiment, we systematically show two figures. The first one describes the time evolution of the individual opinion with respect to time for the whole number of individuals and the second one shows the time evolution of the quantity

$$S_+ = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{x_i(t) > 0},$$

which represents the fraction of individuals which favour the underlying binary question at a given time $t \in \mathbb{R}^+$. Of course, from S_+ is possible to deduce the fraction of individuals which do not approve the underlying binary question at a given time $t \in \mathbb{R}^+$:

$$S_- = 1 - S_+ = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{x_i(t) \leq 0}.$$

When needed, we add the visualization of the interaction matrix and of the evolution of the number of posts.

4.1. **Strongly connected population.** In this first series of tests, the form of the interaction matrix is the following. We first impose that

$$\sigma_{i,i} = 1 \quad \text{for all } i = 1, \dots, N;$$
$$\sigma_{i,i+1} = 1 \quad \text{for all } i = 1, \dots, (N-1) \qquad \sigma_{N,1} = 1.$$
$$\sigma_{i+1,i} = 1 \quad \text{for all } i = 1, \dots, (N-1) \qquad \sigma_{1,N} = 1.$$

These conditions guarantee that the network represented by the interaction matrix is strongly connected. Moreover, we add some extra non-zero entries to the interaction matrix by means of a sampling from the uniform distribution. The explicit form of the interaction matrix is described in Figure 2 (left) and the total number of non-zero entries is equal to 32,439. The initial number of posts of the agents of the population is the following: $b_i^0 = 1$, for all $i = 1, \dots, N$ (see Figure 2, right). The initial condition for the unknowns x_i is:

$$x_i^0 = -0.9999 + 1.9998 \times \frac{i-1}{N-1}, \qquad i = 1, \dots, N. \tag{4.4}$$

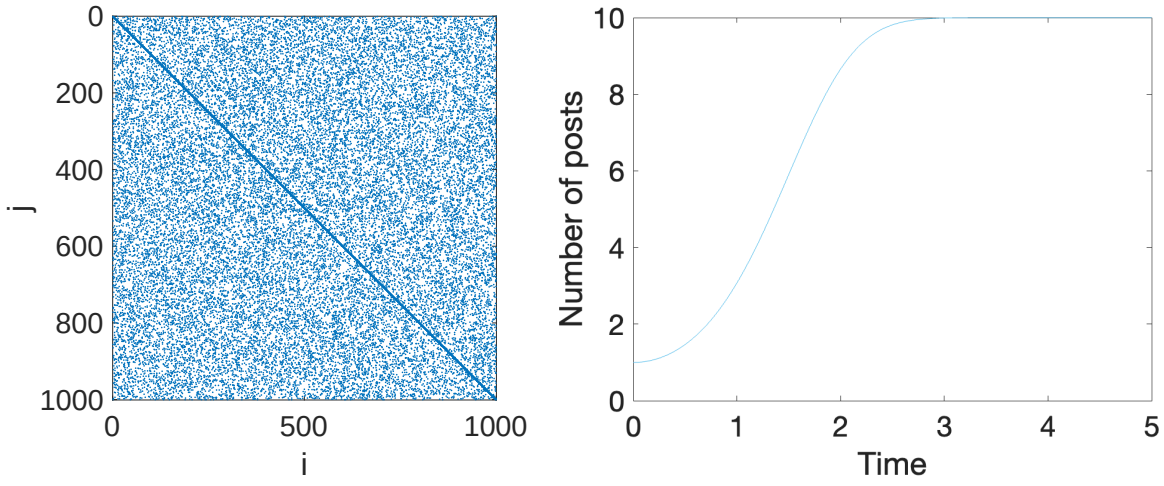

FIGURE 2. Interaction matrix (left) and time evolution of the number of posts (right) in Cases 1 and 2.

4.1.1. *Case 1.* In Case 1, we suppose that the microblogging network's provider is neutral. We see that the geometry of the network representing the connections between the individuals strongly modifies the behaviour of the system. In particular, the system tends to an equilibrium, which is different from zero (see Figure 3, right). This behaviour is the consequence of the non-symmetric interactions in the network representing the interactions between the members of the population. The evolution of the population S_+ starts from 0.5 at time $t = 0$ and reaches in a non-monotone way, at time $t = 5$, the value $S_+(5) = 0.455$. We underline that the high-frequency oscillations in all the graphs of S_+ are not originated by a noise. This effect is purely deterministic and is due to the presence of many opinion trajectories whose ordinate change its sign.

4.1.2. *Case 2.* In this case, we study the effects of hidden manipulation on the same population studied in Case 1. We suppose that the highlighting functions have the form

$$\psi_{i,j}(t, x_i(t), x_j(t)) = \begin{cases} 1 & \text{if } x_j(t) \geq x_i(t) \\ \dfrac{4}{5} & \text{otherwise.} \end{cases}$$

We underline that this form of the highlighting functions drives the system towards positive opinions. This feature of the highlighting functions is confirmed by the numerical experiments. We see that the subpopulation S_+ is weakly oscillating, but reaches in five weeks the value 1,

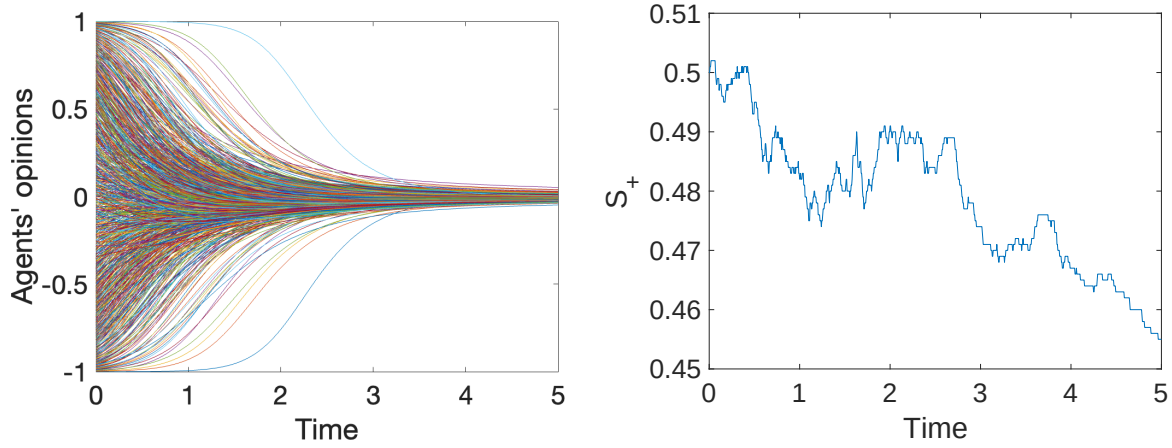

FIGURE 3. Time evolution of the opinions (left) and of S_+ (right) in Case 1.

starting from the value $S_+(0) = 0.5$ (Figure 4, right) and with an interaction matrix which clearly favours the negative opinions, as shown in Case 1. Hence, hidden manipulation is an efficient way for driving the system towards positive opinions.

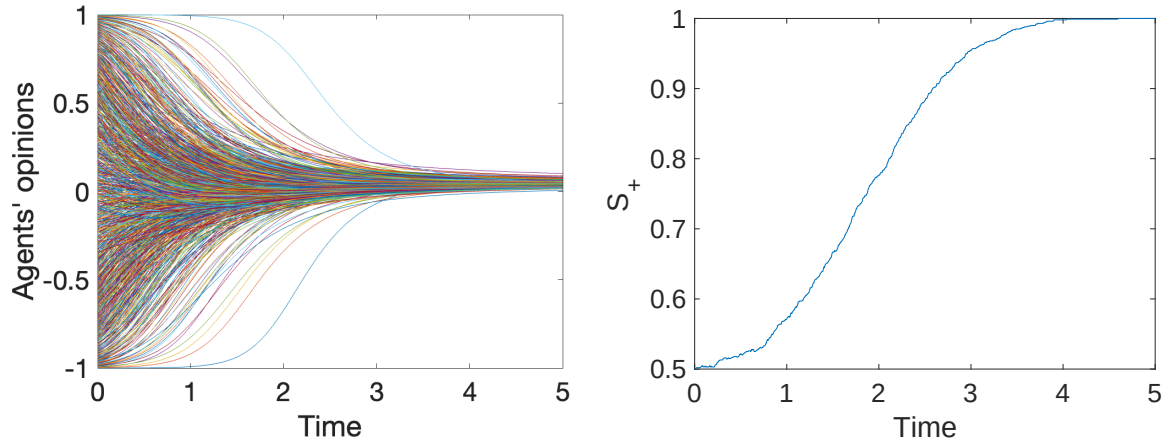

FIGURE 4. Time evolution of the opinions (left) and of S_+ (right) in Case 2.

4.2. **Clustered population.** Another series of tests of this subsection studies a population composed of distinct clusters. These clusters include individuals with different initial viewpoints about the underlying binary question. The initial condition – which is randomly chosen from the uniform distribution – and the interaction matrix are detailed in Figure 5. The total number of connexions of the network is equal to 15,028.

Initially, the population has average opinion equal to -0.0062. The fraction of the population with positive opinion at time $t = 0$ is $S_+(0) = 0.499$. Even if $S_-(0) > S_+(0)$, we see that the average opinion has a major effect on the time evolution of the population, as underlined in Section 2. This indicator can be more important than the fraction of the population having opinions of the same sign.

We will study two types of hidden manipulation. In Case 4, the network highlights opinions which have negative sign whereas Case 5 treats a possible structure of highlighting functions which put in light, for the i-th individual, all the opinions which are closer to -1 than $x_i(t)$.

Being interested in studying the hidden manipulation effects, also in this case we suppose that all the agents have the same blogging activity (i.e. $b_i^0 = 1$ for all $i = 1, \dots, N$).

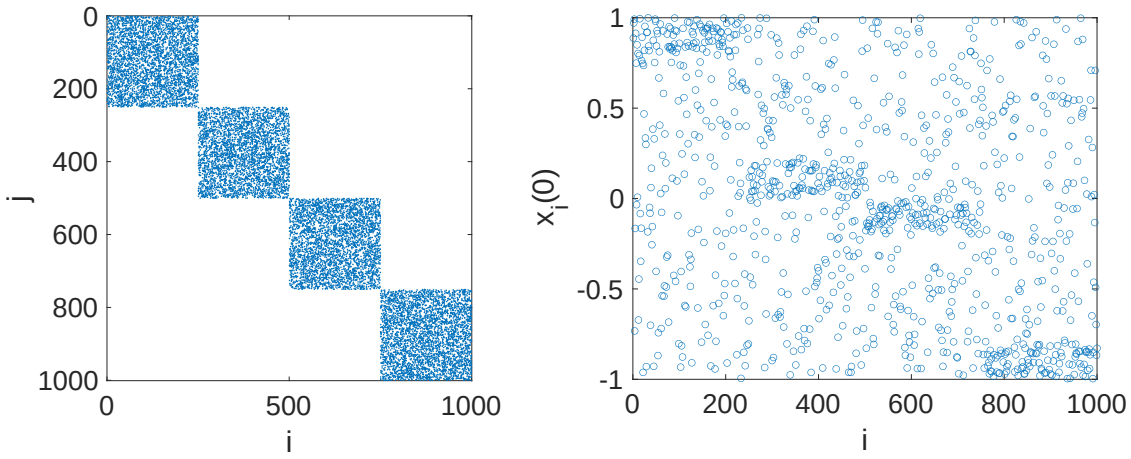

FIGURE 5. Interaction matrix (left) and initial condition (right) in Cases 3, 4 and 5.

4.2.1. *Case 3.* We first treat the situation without hidden manipulation. In Figure 6 we reproduce the individual opinion evolution and the evolution of the fraction of the population with positive opinions. We observe that the agents aggregate themselves in several clusters and that S_+ grows from $S_+(0) = 0.499$ to $S_+(5) = 0.5$. The four detected clusters are consistent with the four interconnected sub-populations of the network.

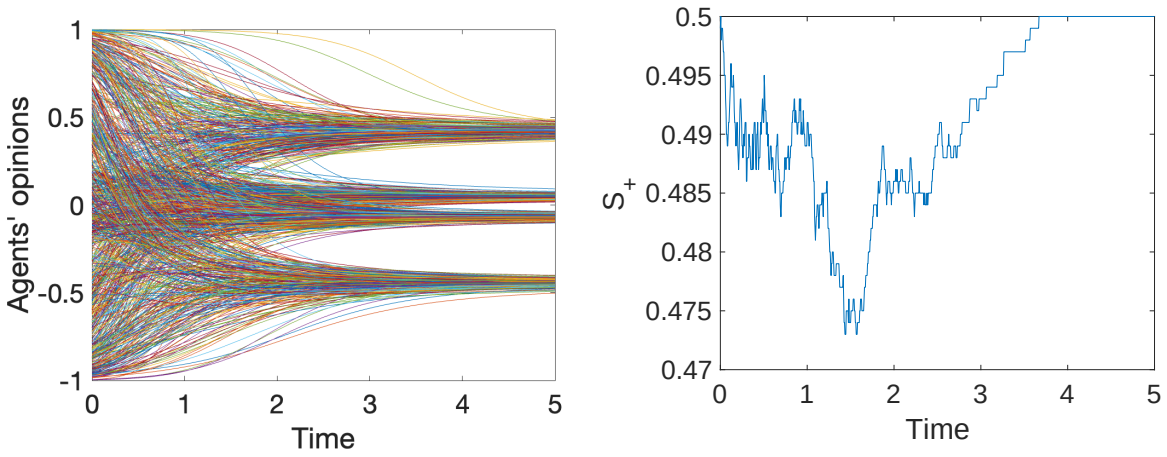

FIGURE 6. Time evolution of the opinions (left) and of S_+ (right) in Case 3.

4.2.2. *Case 4.* The hidden manipulation effect is simulated by using highlighting functions of type

$$\psi_{i,j}(t, x_i(t), x_j(t)) = \begin{cases} 1 & \text{if } x_j(t) \leq 0 \\ \dfrac{4}{5} & \text{otherwise.} \end{cases} \tag{4.5}$$

We underline that this form of the highlighting functions has a decisive effect in pushing the system towards negative opinions. In Figure 7, we note that the subpopulation S_+ decreases from $S_+(0) = 0.499$ to $S_+(5) = 0.320$. Moreover, the whole population reduces itself to four clusters, two of them are centred below zero. The result is very sensitive to the weights of the opinion in the highlighting functions.

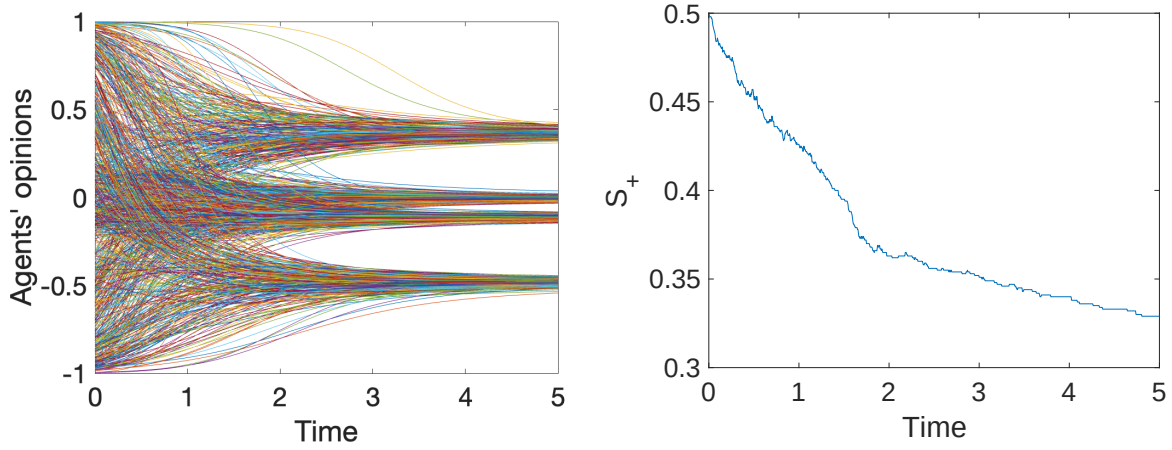

FIGURE 7. Time evolution of the opinions (left) and of S_+ (right) in Case 4.

4.2.3. *Case 5.* The hidden manipulation effect of this simulation is obtained thanks to highlighting functions of type

$$\psi_{i,j}(t, x_i(t), x_j(t)) = \begin{cases} 1 & \text{if } x_i(t) \geq x_j(t) \\ \\ \dfrac{4}{5} & \text{otherwise.} \end{cases}$$

Figure 8 shows that this strategy is less efficient than the strategy used in Case 4, even if this strategy could help in decreasing the value of the opinion variable of individuals which are exclusively in contact with individuals of positive opinion. However, this strategy is enough for driving the system to $S_+(5) = 0.486$.
In order to compare with greater accuracy the behaviour of S_+ in these different cases, figure 9 shows on the same scale the trend of this variable.

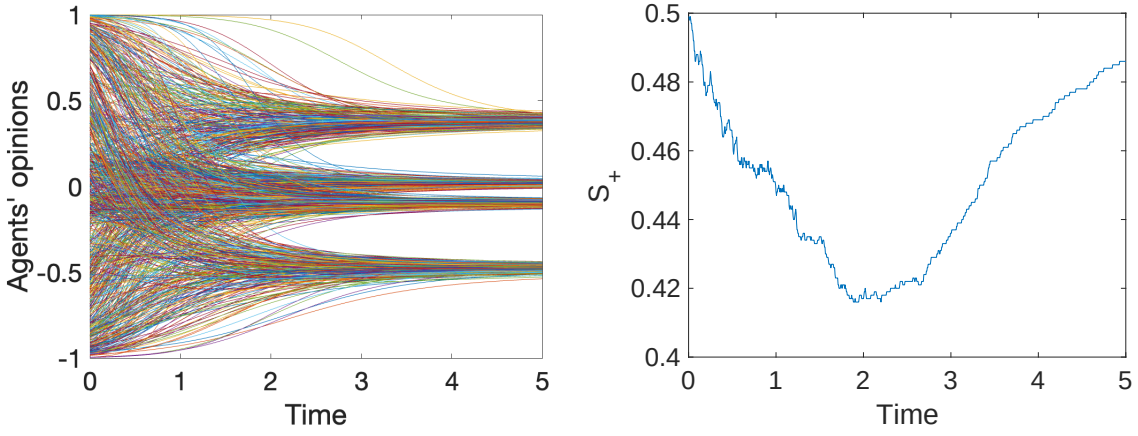

FIGURE 8. Time evolution of the opinions (left) and of S_+ (right) in Case 5.

4.3. **Effects of a leadership.** In the previous scenarios, we studied the process of opinion formation in the presence or absence of hidden manipulation. Another way to influence the opinions of a population is the presence of one or more leaders within the community. A leader is, for our purposes, an agent who interact with the population, is not influenced by it but attracts the opinions of other individuals towards his/her positions. Therefore, it can be interesting to compare the effects of hidden manipulation and the presence of a leader in a network. Three situations will then be compared: the first one is a neutral situation, in

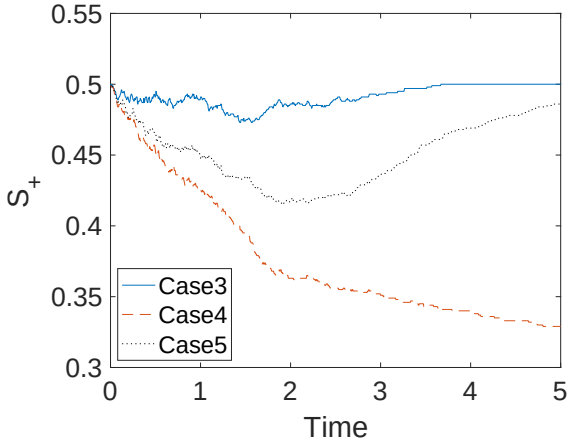

FIGURE 9. Comparison of the behaviour of S_+ in Case 3, 4 and 5.

the second one hidden manipulation is present, and, in the third ones, a leader influences the population. The structure of the model is the same adopted in the previous sections. In all simulations, we consider a population composed of $N = 10^3$ agents. The interaction matrix is composed of five distinct clusters with 31,457 non-zero entries. Opinions are initially distributed as follows:

$$z_i = -0.9999 + 1.9998 \times \frac{i-1}{N-1}, \qquad i = 1, \dots, N. \tag{4.6}$$

We then have:

$$x_i^0 = z_i|z_i|, \qquad i = 1, \dots, N. \tag{4.7}$$

Figure 10 shows the interaction matrix and the initial opinions.

In this series of tests, individuals interact with a bounded confidence level. This term is included in the highlighting functions. The confidence level is equal to 0.6. The set of parameters considered is the following: $\beta = 2$, $\gamma = 10$, $\mu = 10^{-4}$, $\Delta t = h = 5 \times 10^{-3}$ and the simulations have been displayed for $t \in (0, 100)$.

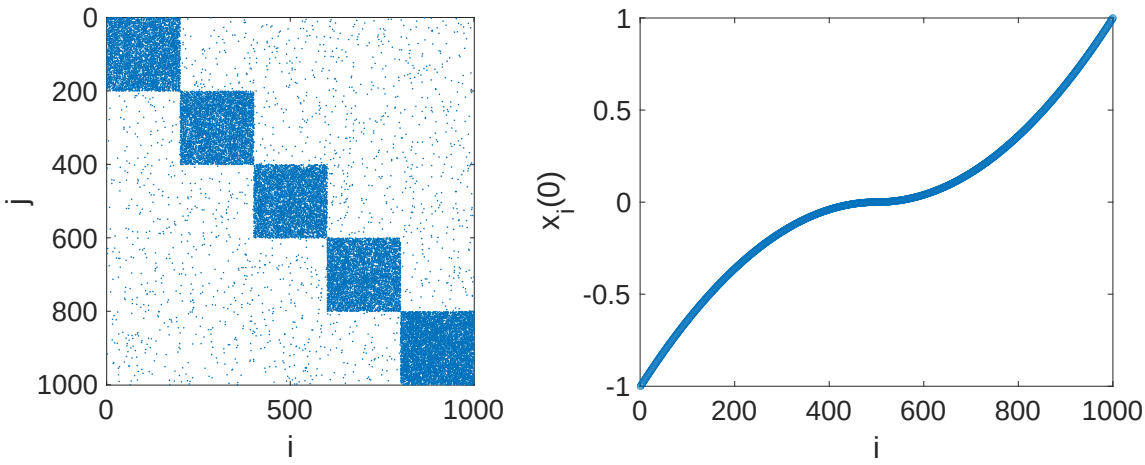

FIGURE 10. Interaction matrix (left) and initial conditions (right)

4.3.1. *Neutral network.* We have, at first, studied the behaviour of the opinion formation in absence of hidden manipulation and without any leader. The evolution of the population S_+ starts from 0.5 at time $t = 0$ and reaches in a non-monotone way the value $S_+ = 0.6$ at time $t = 100$.

4.3.2. *Hidden manipulation.* We have then studied the evolution of the opinions in presence of hidden manipulation. The hidden manipulation effect of this simulation is obtained thanks to highlighting functions of type

$$\psi_{i,j}(t, x_i(t), x_j(t)) = \begin{cases} \mathbb{1}_{|x_i(t)-x_j(t)|<3/5} & \text{if } x_i(t) \geq x_j(t) \\ \frac{1}{2}\mathbb{1}_{|x_i(t)-x_j(t)|<3/5} & \text{otherwise.} \end{cases}$$

We underline that $S_+ = 0.565$ at time $t = 100$, and this is in line with the form of the highlighting functions.

4.3.3. *Presence of a leader.* The last scenario studies the effect of a leader on the opinion of the population, in the absence of hidden manipulation. For consistency with the other tests, in this case $N = 1.001 \times 10^3$ and the leader is denoted by the index N. The leader has initial opinion $x_N = -0.3$. The highlighting functions involving the leader have much greater weight than the highlighting functions of the other agents. Specifically, we determined that each member of the population has weight equal to 1, while the leader has weight equal to 10^4. Moreover, the leader is not influenced by the other agents of the population:

$$\psi_{i,j}(t, x_i(t), x_j(t)) = \begin{cases} \mathbb{1}_{|x_i(t)-x_j(t)|<3/5} & \text{if } i = 1, \dots, N-1 \text{ and } j = 1, \dots, N-1 \\ 10^4 \times \mathbb{1}_{|x_i(t)-x_j(t)|<3/5} & \text{if } i = 1, \dots, N-1 \text{ and } j = N \\ 0 & \text{if } i = N \text{ and } j = 1, \dots, N-1. \end{cases}$$

The simulation shows that $S_+ = 594$ at $t = 100$, as indicated in Figure 11. Therefore, even though the leader influences the opinion of members of the population, in the situations examined in this paper, hidden manipulation is more efficient in changing opinions within the community. To better compare the behaviour of S_+, we report all three results in Figure 11.

The behaviour does not vary very much if the weight of the leader's opinion has the value 10^6.

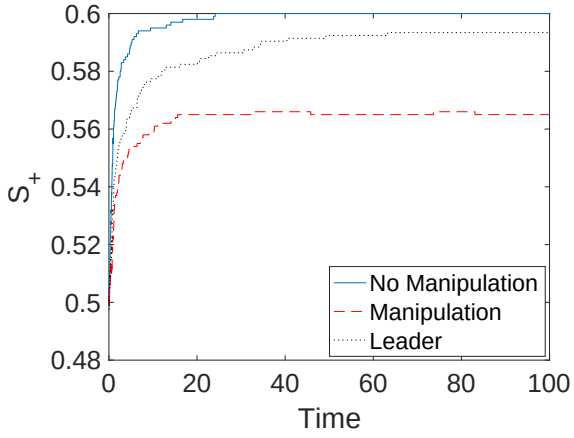

FIGURE 11. Time history of the quantity S_+ for the cases described in Subsections 4.3.1, 4.3.2 and 4.3.3.

4.4. **Stability analysis.** In this subsection, we study the robustness of the outputs of the model with respect to the form of the interaction matrix (which describes the degree of connectivity of the network). All the simulations have been obtained for $N = 100$ agents and by considering the initial condition

$$\begin{cases} x_i(0) = -1 + \frac{7}{10}(i-1)\frac{2}{N-2} & \text{if } i = 1, \dots, \frac{N}{2} \\ x_i(0) = 1 - \frac{7}{10}(N-i)\frac{2}{N-2} & \text{if } i = \frac{N}{2}+1, \dots, N. \end{cases}$$

As a consequence, the initial fraction of the population with negative opinions has exactly the same magnitude as the fraction of the population with positive opinions. In all the simulations, we have used the highlighting function family (4.5).

We have studied the evolution of the population by supposing to modify, at each run, the geometry of the network (or, in other words, the values of the interaction matrix), by imposing an upper bound on the number of non-zero elements of the interaction matrix. Then, we have studied the efficacy of the highlighting function family in driving the opinions of the population towards negative opinions.

Let $N_{\max}$ be the maximal number of non-zero elements of the interaction matrix (which is obviously linked to the sparsity index of the matrix).

We have considered the following values of $N_{\max}$: $N_{\max} = 600$, $N_{\max} = 700$, $N_{\max} = 10^3$, $N_{\max} = 1,4 \times 10^3$, $N_{\max} = 2 \times 10^3$, $N_{\max} = 2.5 \times 10^3$, $N_{\max} = 3 \times 10^3$. For each value of $N_{\max}$, we have studied 10 numerical simulations, each of them with a different interaction matrix. As far as the sparsity index increases, the number of simulations which have a majority of positive opinions at time $t = 5$ decreases and, consequently, the quantity $S_+(5)$ decreases.

In two tables, we have collected some relevant results.

Label	Number of non-zero $\sigma_{i,j}$	$S_-(5)$ without manipulation	$S_+(5)$ without manipulation	$S_-(5)$ with manipulation	$S_+(5)$ with manipulation
1	672	0.85	0.15	0.96	0.04
2	671	0.19	0.81	0.92	0.08
3	677	0.08	0.92	0.87	0.13
4	671	0.98	0.02	0.98	0.02
5	677	0.30	0.70	0.94	0.06
6	683	0.04	0.96	0.21	0.79
7	673	0.13	0.87	0.92	0.08
8	684	0.15	0.85	0.92	0.08
9	676	0.66	0.34	0.92	0.08
10	679	0.90	0.10	0.97	0.03

TABLE 1. Comparison between 10 simulations with random entries of the interaction matrix ($N_{\max} = 700$).

In Table 1, we have labelled, in the first column, the number of the simulation and then we have collected, in the second column the number of non-zero entries of the matrix representing the network and, in the third and fourth column, the values of $S_-(5)$ and $S_+(5)$ without hidden manipulation effect. As expected, the population reaches an equilibrium in which the numbers of positive and negative opinions heavily depend on the connexions between the agents, represented by the non-zero entries of the interaction matrix.

However, if the system is under the effect of the highlighting function family (4.5), which favours the negative opinion, the behaviour of the population is clearly influenced by it. Its quantification is collected in the fifth and in the sixth column. We observe that the highlighting functions family has always a non-negligible effect and that, in some cases, it is able to renverse the majority inside the population.

For the sake of completeness, we have collected in Table 2 the results in the case $N_{\max} = 1.4 \times 10^3$. At each run, we have considered a different strongly connected network obeying to the same hypotheses as in Subsection 4.1, and by varying the value of $N_{\max}$. It is apparent that, when the population is sufficiently interconnected, the manipulation effect of the highlighting functions becomes more efficient and, above a given threshold, it is the dominant effect.

The simulations suggest hence that there is a threshold effect on the sparsity index of the interaction matrix.

Label	Number of non-zero $\sigma_{i,j}$	$S_-(5)$ without manipulation	$S_+(5)$ without manipulation	$S_-(5)$ with manipulation	$S_+(5)$ with manipulation
1	1377	0.12	0.88	0.92	0.08
2	1395	0.17	0.83	0.93	0.07
3	1387	0.88	0.12	0.99	0.01
4	1393	0.14	0.86	0.91	0.09
5	1395	0.73	0.27	0.99	0.01
6	1396	0.16	0.84	0.92	0.08
7	1391	0.52	0.48	0.99	0.01
8	1391	0.89	0.11	0.99	0.01
9	1399	0.23	0.77	0.95	0.05
10	1386	0.36	0.64	0.97	0.03

TABLE 2. Comparison between 10 simulations with random entries of the interaction matrix ($N_{\max} = 1.4 \times 10^3$).

5. Conclusion

We have studied some dynamics of opinion dynamics. In the simulations, we have considered a fixed network, but the model allows to treat, in the same way, an evolutionary network. The model takes into account the effects on public opinion caused by the sign and the intensity of the initial opinions of the agents, their activity in microblogging platforms and the possible manipulations of the visibility of the posts by the microblogging platform provider. We have numerically studied the stability of the hidden manipulation phenomenon with respect to the sparsity of the network and we have shown that very mild interventions of the network owner can have major effects on the opinion of the population.

Moreover, we have shown simulations suggesting that, in some situations, hidden manipulation is more efficient than a leader in modifying the opinion of a population, at least if the time interval is not too large.

Hence, hidden manipulation may have an important impact on the public opinion formation.

Appendix

In several situations, it may be convenient to work in discrete-time, especially if it is necessary to consider punctual events. For this reason, we describe here the discrete-time version of our model. Let $i = 1, \dots, N$ the label of each agent and let $t \in \mathbb{N}$. Let $b_i : \mathbb{N} \to \mathbb{R}^+$ the number of posts sent to the network by the i-th agent, $\sigma_{i,j} : \mathbb{N} \to \{0, 1\}$ for all $i,\ j = 1, \dots, N$ the entries of the discrete-time interaction matrix, and $x_i : \mathbb{N} \to [-1, 1]$ the opinion of the i-th agent. The discrete-time model has the following form:

$$\begin{cases} b_i(t+1) = b_i(t) + \mu_i(\gamma_i - b_i(t)) \displaystyle\sum_{j=1}^{N} \sigma_{i,j}(t) \sum_{k=0}^{t} b_j(k) \qquad \gamma_i,\ \mu_i > 0 \\ \\ x_i(t+1) = x_i(t) + \alpha_i(x_i(t))(\Phi_i(t) - x_i(t)), \end{cases} \tag{5.1}$$

where

$$\Phi_i(t) = \begin{cases} \dfrac{\sum_{j=1}^{N} \sigma_{i,j}(t) \sum_{k=0}^{t} b_j(k) x_j(k) \psi_{i,j}(k)}{\sum_{j=1}^{N} \sigma_{i,j}(t) \sum_{k=0}^{t} b_j(k) \psi_{i,j}(k)} & t > 0 \\ \dfrac{\sum_{j=1}^{N} \sigma_{i,j}(0) b_j(0) x_j(0) \psi_{i,j}(0)}{\sum_{j=1}^{N} \sigma_{i,j}(0) b_j(0) \psi_{i,j}(0)} & t = 0 \end{cases} \tag{5.2}$$

and

$$\psi_{i,j} = \begin{cases} \psi_{i,j}(k, x_j(k), b_j(k)) > 0 & i \neq j \\ 1 & i = j. \end{cases} \tag{5.3}$$

The model is coupled with suitable initial conditions: for all $i = 1, \dots, N$

$$b_i(0) = b_i^0 \in (0, \gamma_i), \tag{5.4}$$

$$x_i(0) = x_i^0 \in \Omega. \tag{5.5}$$

We underline that the discrete-time Equations (5.1)-(5.5) have the same structure as the first-order Euler scheme of the continuous-time model (2.1)-(2.5), with time step equal to one.

The model can be further simplified if we suppose that the number of posts published by each agent is constant with respect to time, i.e. the posts production has rapidly reached the saturation regime. Under this assumption, if we denote with γ_i the number of posts written by the i-th agent at each time step, the discrete-time model has the following structure:

$$x_i(t+1) = x_i(t) + \alpha_i(x_i(t))(\Phi_i(t) - x_i(t)),$$

where

$$\Phi_i(t) = \begin{cases} \dfrac{\sum_{j=1}^{N} \sigma_{i,j}(t) \sum_{k=0}^{t} \gamma_j x_j(k) \psi_{i,j}(k)}{\sum_{j=1}^{N} \sigma_{i,j}(t) \sum_{k=0}^{t} \gamma_j \psi_{i,j}(k)} & t > 0 \\ \dfrac{\sum_{j=1}^{N} \sigma_{i,j}(0) \gamma_j(0) x_j(0) \psi_{i,j}(0)}{\sum_{j=1}^{N} \sigma_{i,j}(0) \gamma_j \psi_{i,j}(0)} & t = 0 \end{cases}$$

and

$$\psi_{i,j} = \begin{cases} \psi_{i,j}(k, x_j(k), \gamma_j) > 0 & i \neq j \\ 1 & i = j. \end{cases}$$

The initial conditions are $x_i(0) = x_i^0 \in \Omega$ for all $i = 1, \dots, N$.

Acknowledgements. This work has been carried out in the framework of the project *Kimega* (ANR-14-ACHN-0030-01). This research was moreover supported by the Italian Ministry of Education, University and Research (MIUR), *Dipartimenti di Eccellenza* Program - Department of Mathematics "F. Casorati", University of Pavia. The authors tank the anonymous referees for their useful comments and suggestions.

References

[1] Robert P Abelson. Mathematical models of the distribution of attitudes under controversy. *Contributions to mathematical psychology*, 1964.

[2] Franz Achleitner, Anton Arnold, and Eric A. Carlen. On multi-dimensional hypocoercive BGK models. *Kinet. Relat. Models*, 11(4):953–1009, 2018.

[3] Mohammed N. Al-Rashdan, Malak Abdullah, Mahmoud Al-Ayyoub, and Yaser Jararweh. Authorship analysis of English and Spanish tweets. *Proceedings of the Association for Information Science and Technology*, 57(1):e261, 2020.

[4] G. Albi, L. Pareschi, and M. Zanella. Boltzmann-type control of opinion consensus through leaders. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 372(2028):20140138, 18, 2014.

[5] Luigi Ambrosio. Well posedness of ODE's and continuity equations with nonsmooth vector fields, and applications. *Rev. Mat. Complut.*, 30(3):427–450, 2017.

[6] Sven Banisch, Ricardo Lima, and Tanya Araújo. Agent based models and opinion dynamics as Markov chains. *Social Networks*, 34(4):549–561, 2012.

[7] Laurent Boudin, Aurore Mercier, and Francesco Salvarani. Conciliatory and contradictory dynamics in opinion formation. *Physica A: Statistical Mechanics and its Applications*, 391(22):5672 – 5684, 2012.

[8] Laurent Boudin, Roberto Monaco, and Francesco Salvarani. Kinetic model for multidimensional opinion formation. *Phys. Rev. E (3)*, 81(3):036109, 9, 2010.

[9] Laurent Boudin and Francesco Salvarani. A kinetic approach to the study of opinion formation. *M2AN Math. Model. Numer. Anal.*, 43(3):507–522, 2009.

[10] Laurent Boudin and Francesco Salvarani. The quasi-invariant limit for a kinetic model of sociological collective behavior. *Kinet. Relat. Models*, 2(3):433–449, 2009.

[11] Laurent Boudin and Francesco Salvarani. Opinion dynamics: Kinetic modelling with mass media, application to the Scottish independence referendum. *Phys. A*, 444:448–457, 2016.

[12] Laurent Boudin, Francesco Salvarani, and Emmanuel Trélat. Exponential convergence towards consensus for non-symmetric linear first-order systems in finite and infinite dimensions, 2021. arXiv:2104.14183.

[13] Engin Bozdag. Bias in algorithmic filtering and personalization. *Ethics and information technology*, 15(3):209–227, 2013.

[14] Marilynn B Brewer. The social self: On being the same and different at the same time. *Personality and social psychology bulletin*, 17(5):475–482, 1991.

[15] Francesca Ceragioli and Paolo Frasca. Continuous and discontinuous opinion dynamics with bounded confidence. *Nonlinear Anal. Real World Appl.*, 13(3):1239–1251, 2012.

[16] Felipe Cucker and Steve Smale. Emergent behavior in flocks. *IEEE Trans. Automat. Control*, 52(5):852–862, 2007.

[17] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch. Mixing beliefs among interacting agents. *Adv. Complex Systems*, 3:87–98, 2000.

[18] Anneliese Depoux, Sam Martin, Emilie Karafillakis, Raman Preet, Annelies Wilder-Smith, and Heidi Larson. The pandemic of social media panic travels faster than the COVID-19 outbreak. *Journal of Travel Medicine*, 27(3), 03 2020.

[19] Elizabeth Dubois and Devin Gaffney. The multiple facets of influence: Identifying political influentials and opinion leaders on twitter. *American behavioral scientist*, 58(10):1260–1277, 2014.

[20] Andreas Flache, Michael Mäs, Thomas Feliciani, Edmund Chattoe-Brown, Guillaume Deffuant, Sylvie Huet, and Jan Lorenz. Models of social influence: Towards the next frontiers. *Journal of Artificial Societies and Social Simulation*, 20(4):2, 2017.

[21] Daniel Geschke, Jan Lorenz, and Peter Holtz. The triple-filter bubble: Using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers. *British Journal of Social Psychology*, 58(1):129–149, 2019.

[22] Petri Hallikainen. Why people use social media platforms: Exploring the motivations and consequences of use. In Lapo Mola, Ferdinando Pennarola, and Stefano Za, editors, *From Information to Smart Society*, pages 9–17, Cham, 2015. Springer International Publishing.

[23] R. Hegselmann and U. Krause. Opinion dynamics and bounded confidence: models, analysis and simulation. *J. Artif. Soc. Soc. Sim.*, 5(3), 2002.

[24] Rainer Hegselmann and Ulrich Krause. Opinion dynamics under the influence of radical groups, charismatic leaders, and other constant signals: A simple unifying model. *Networks & Heterogeneous Media*, 10(3):477, 2015.

[25] Clinton Innes, Razvan C Fetecau, and Ralf W Wittenberg. Modelling heterogeneity and an open-mindedness social norm in opinion dynamics. *Networks & Heterogeneous Media*, 12(1):59, 2017.

[26] J. Isaak and M. J. Hanna. User data privacy: Facebook, Cambridge Analytica, and privacy protection. *Computer*, 51(8):56–59, 2018.

[27] Michael Klenk. (Online) manipulation: sometimes hidden, always careless. *Review of Social Economy*, 0(0):1–21, 2021.

[28] L. Li, Q. Zhang, X. Wang, J. Zhang, T. Wang, T. Gao, W. Duan, K. K. Tsoi, and F. Wang. Characterizing the propagation of situational information in social media during COVID-19 epidemic: A case study on Weibo. *IEEE Transactions on Computational Social Systems*, 7(2):556–562, 2020.
[29] Loretta Mastroeni, Pierluigi Vellucci, and Maurizio Naldi. Agent-based models for opinion formation: A bibliographic survey. *IEEE Access*, 7:58836–58848, 2019.
[30] Rosa Maria Paniccia. How the internet changes in the time of the covid19 pandemic. *Rivista di Psicologia Clinica*, 15(1):29–46, 2020.
[31] Eli Pariser. *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011.
[32] Talcott Parsons, Edward A Shils, et al. Values, motives, and systems of action. *Toward a general theory of action*, 33:247–275, 1951.
[33] Manuel Gomez Rodriguez, Krishna Gummadi, and Bernhard Schoelkopf. Quantifying information overload in social media and its impact on social contagions. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.
[34] Thomas J Scheff. Toward a sociological model of consensus. *American Sociological Review*, pages 32–46, 1967.
[35] Giuseppe Toscani. Kinetic models of opinion formation. *Commun. Math. Sci.*, 4(3):481–496, 2006.
[36] Wenxian Wang, Xingshu Chen, Shuyu Jiang, Haizhou Wang, Mingyong Yin, and Peiming Wang. Exploring the construction and infiltration strategies of social bots in Sina microblog. *Scientific Reports*, 10(1):1–19, 2020.

(G. B.) Istituto di Istruzione Superiore "Alberti", Via Monte Confinale 10, 23032 Bormio, Italy

(F. S.) Léonard de Vinci Pôle Universitaire, Research Center, 92916 Paris La Défense, France & Dipartimento di Matematica "F. Casorati", Università degli Studi di Pavia, Via Ferrata 1, 27100 Pavia, Italy

Email address: giulia.braghini.1995@gmail.com
Email address: francesco.salvarani@unipv.it