ON THE DETECTION OF HIDDEN OPINION MANIPULATION IN MICROBLOGGING PLATFORMS

Giulia Braghini, Francesco Salvarani

ON THE DETECTION OF HIDDEN OPINION MANIPULATION IN MICROBLOGGING PLATFORMS

GIULIA BRAGHINI AND FRANCESCO SALVARANI

Abstract. This article studies a new model for describing opinion manipulation effects on a microblogging directed network system. The network can evolve in time, by means of the creation and the deletion of some connexions. 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 and the possible manipulations of the visibility of the posts by the microblogging platform owner. 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. We moreover suggest a strategy for detecting whether the blogging platform is neutral or not.

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. For this reason, 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.

On the other hand, the recent COVID-19 disease has caused an explosion of web-based contents related to the pandemic. The intrinsic interest of the subject and the social distancing due to the prevention health policies have given an unexpected contribution to the online dissemination of verified or unverified contents [6, 9].

The authors of web-based contents may have several reasons which explain their actions. Many writers 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.

In many cases, the influencers are known, but sometimes they are not. The explicit strategy is the easiest to be detected: the publisher of a content explicitly declares his viewpoint and develops all the arguments which support his thesis. However, when the public opinion is driven by means of non-declared manipulation techniques, many individuals may be the object of hidden manipulation, with possible serious consequences for the collective behaviour in democratic societies.

A very well-known case is the so-called *Facebook-Cambridge Analytica* data scandal [8], which highlighted some possible consequences of social media sharing, especially in connexion with non-apparent opinion manipulation techniques related to elections or referendums.

The goal of this article is to understand and to contrast the effect of hidden influencing through social media. We focus on microblogging and social networking services on which users may post messages and interact with other users. The sociological literature agrees that one of the main dynamics of opinion formation is *consensus* [10, 12] and, consequently, many

Date: August 31, 2020.

mathematical models of social interaction, such as the well-known Hegselmann-Krause model [7] and the Cucker-Smale model [5] are based on consensus dynamics.

Opinion formation being mainly driven by consensus, it is clear that the neutrality of the blogging platform is crucial for democracy. Indeed, if the platform is not neutral and ranks the posted opinions in a deliberated way, then the process of opinion formation is biased.

In this article we focus on 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, because it works with trusted posts of real individuals, with no doubt about the authors, which are freely chosen by the follower, and which can influence him.

This technique requires that the owner of the microblogging platform network is involved in the manipulation dynamics.

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 policy of the blogging platform in visualizing the posts among the interested users. We study how the platform policy may modify the public opinion and we compare the results with the unbiased situation, in which all the opinions are treated in an equal way.

The structure of the article is the following. In Section 2, we describe our mathematical model. We discuss some mathematical questions on the model in Section 3 and then we implement the model, test it on some relevant cases, and analyse its main consequences in Section 4. Finally, we conclude our analysis by suggesting a strategy for detecting such type of hidden manipulation.

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 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:

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

where

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

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 up to a saturation phenomenon, with saturation constant $\gamma_i > 0$. 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 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 own opinion through a consensus dynamics by taking into account the opinions of the individuals he is following. In (2.1), the time evolution of the functions x_i is governed by the joint contribution of two terms. The functions $\alpha_i : \Omega \to \mathbb{R}^+$ is somehow the analogous of the admissible functions defined in [3] (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 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 describe the possible manipulation induced by the owner of the platform. These quantities are then 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 has a complete access to all his posts, and, when a user decide to eliminate another individual from his set of contacts, he looses the access to all his comments.

As said before, the highlighting functions $\psi_{i,j}(t)$ describe the manipulation effect. These terms are supposed to be under the control of the microblogging network's provider. When there is no manipulation, $\psi_{i,j}(t) = 1$ for all $i, j = 1, \dots, N$. In the case of hidden manipulation, we have that $\psi_{i,i}(t) = 1$ for all $i = 1, \dots, N$ for preventing the individual to see any manipulation effect on his 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 $i, j = 1, \dots, N$). We underline that 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. Timing is a crucial factor in opinion formation dynamics (see [2] for a discussion about this point): for this reason we allow that the manipulation strategy may vary in time. 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 [11]: 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.

Note that, if $\psi_{i,j} = 1$ 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 [7].

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, [4]).

3. Basic mathematical properties of the model

This section is devoted to the mathematical analysis of our model. In particular, we will discuss the existence and uniqueness of the solution, which are two key features of any good mathematical model. 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.,

$$\begin{cases} \dfrac{db_i}{dt}(t) = F_i(t, b_1, \dots, b_N(t)) := \mu_i(\gamma_i - b_i(t)) \displaystyle\sum_{j=1}^{N} \sigma_{i,j}(t) \int_0^t b_j(\theta)\, d\theta \\ \\ b_i(0) = b_i^0 \in (0, \gamma) \end{cases} \tag{3.1}$$

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 (3.1) with initial data $b_i(0) = b_i^0$.

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

$$\begin{cases} \dfrac{dx_i}{dt}(t) = \alpha_i(x_i(t))(\Phi_i(t) - x_i(t)) \qquad i = 1, \dots, N \\ \\ x_i(0) = x_i^0 \in \Omega. \end{cases} \tag{3.2}$$

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 [1] and to the references therein for a complete introduction to the theory of flows associated to non-smooth vector fields).

The Cauchy problem (3.2) has the following structure:

$$\begin{cases} \dot{x}(t) = \eta(t, x(t)) \\ x(0) = x_0, \end{cases} \tag{3.3}$$

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 [1]):

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 [1]:

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* (3.3).

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 (3.1) 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 (3.2). Both systems have been discretized by means of a standard fourth-order Runge-Kutta routine.

In what follows, we always suppose that

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

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 three network geometries. The first geometry describes a fully connected system; the second one considers a strongly connected network (i.e. there exists a path linking each pair of agents of the population) and the third 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 three situations: the first one without hidden manipulation, and the other two with two types of hidden manipulation, which may be active for the whole duration of the simulation or not.

In all the numerical simulations, we consider a population composed of $N = 100$ totally or partially interconnected individuals. The choice of this value for N allows to produce readable figures. 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. We moreover choose the following numerical values: for all i, $\mu_i = \mu^* = 10^{-4}$ (relaxation constant for the first set of equations),

$\gamma_i = \gamma^* = 15$ (maximum number of daily posts for each agent) and $\beta = 2$ (relaxation constant for the second set of equations). The time step of the Runge-Kutta algorithm is $\Delta t = 5 \times 10^{-4}$ and the simulations have been displayed for $t \in [0, 5]$, t being measured in weeks.

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. **Fully interconnected population.** In this first series of cases, we suppose that $\sigma_{i,j} = 1$ for all $i, j = 1, \dots, N$. At the individual level, we suppose that all the agents of the population have identical features: they post the same number of comments on the social network (the initial condition is $b_i^0 = 1$ for all $i = 1, \dots, N$), and differ only with respect to their initial opinion. Figure 1 (left) describes the interaction matrix, whose entries are all equal to one. Figure 1 (right) summarizes the time evolution of the number of posts of each agent. Because of the initial conditions, all the curves are superposed.

The initial condition for the opinion functions is the following:

$$x_i^0 = -0.9999 + 1.9998 \times \frac{i-1}{N-1}, \qquad i = 1, \dots, N. \tag{4.3}$$

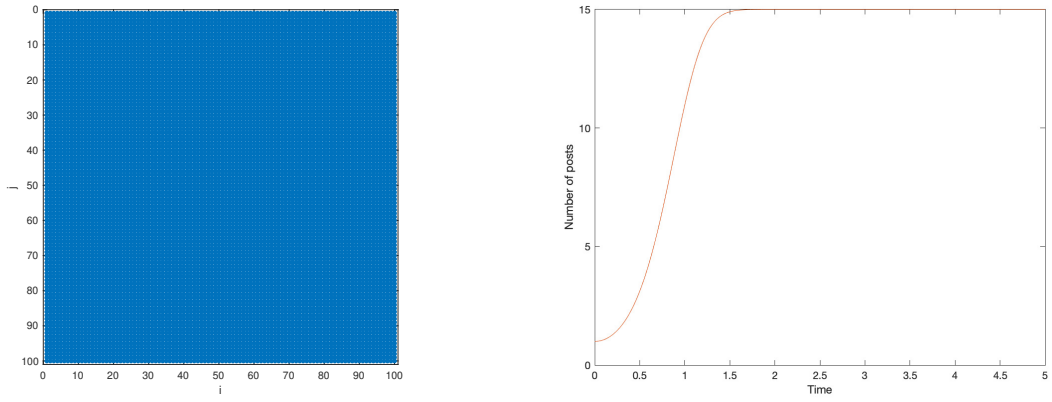

FIGURE 1. Interaction matrix and time evolution of the number of posts (Cases 1, 2 and 3).

4.1.1. *Case 1.* The results of the first simulation, without hidden manipulation (i.e., $\psi_{i,j} = 1$ for all $i, j = 1, \dots, N$), are shown in Figure 2. The average opinion is constant in time and the system relaxes to the equilibrium $x_\infty = 0$ (as expected because of the symmetry of the problem); moreover, individuals with extreme opinions tend more slowly to the equilibrium than individuals with opinions close to the average (see Figure 2, left). The symmetry of the problem and of the initial conditions, together with the other conditions chosen for this test, lead to a perfect equilibrium between the subpopulation favorable to the underlying question and the fraction of the population which is against the underlying question.

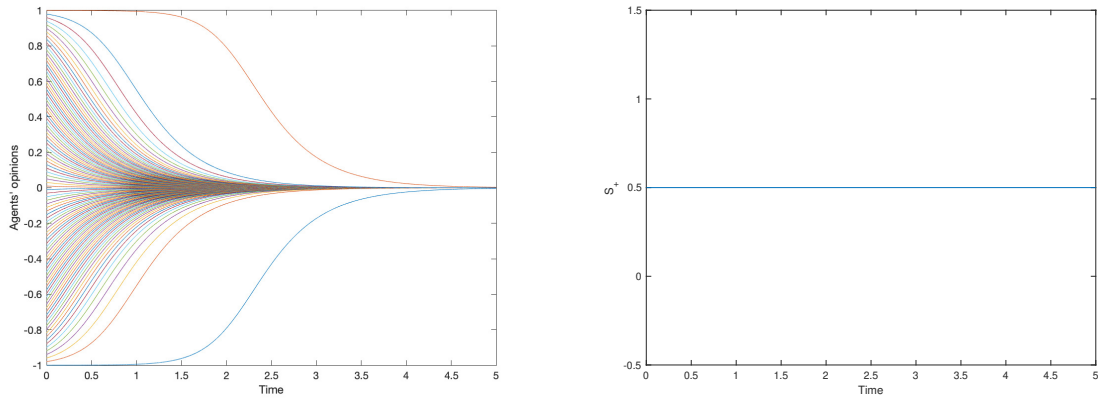

FIGURE 2. Time evolution of the opinions and of S_+ (Case 1).

4.1.2. *Case 2.* We introduce an opinion manipulation technique described by the highlighting functions

$$\psi_{i,j}(t, x_i(t), x_j(t)) = \begin{cases} 1 & \text{if } x_i(t) \geq x_j(t) \\ \\ \dfrac{1}{2} & \text{otherwise.} \end{cases}$$

This choice induces a dynamics towards negative opinions. The other settings stay unchanged with respect to Case 1. In Figure 3 and 4 we show the corresponding numerical results on the time evolution of the opinions and on the time history of S_+. Note that the interaction matrix and time evolution of the number of posts are the same as in Case 1 (see Figure 1).

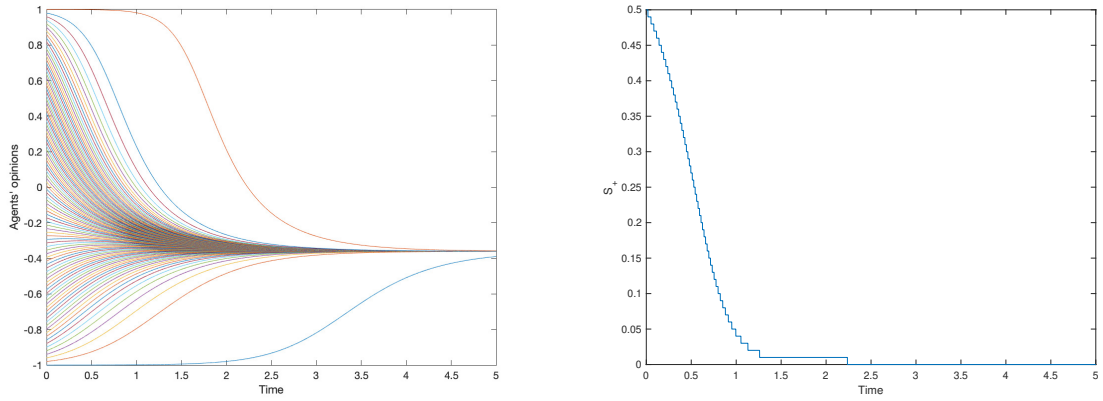

FIGURE 3. Time evolution of the opinions and of S_+ (Case 2).

4.1.3. *Case 3.* The opinion manipulation technique is described, in this scenario, by the highlighting functions

$$\psi_{i,j}(t, x_i(t), x_j(t)) = \begin{cases} 1 & \text{if } x_i(t) \geq x_j(t) \text{ or } t \leq 4 \\ \\ \dfrac{9}{10} & \text{otherwise.} \end{cases}$$

We underline that $\psi_{i,j} = 1$ for all i and for all j when $t \leq 4$. In practice, this family of highlighting functions is a mixing of the previous choice, and induces a dynamics towards negative opinions. The other settings stay unchanged with respect to Case 1. In Figure 4 we show the corresponding numerical results on the time evolution of the opinions and on the time history of S_+. The interaction matrix and time evolution of the number of posts are the

same as in Case 1 (see Figure 1). The presence of a mild manipulation, active only in the last week, induces a non-negligible modification of the population S_+.

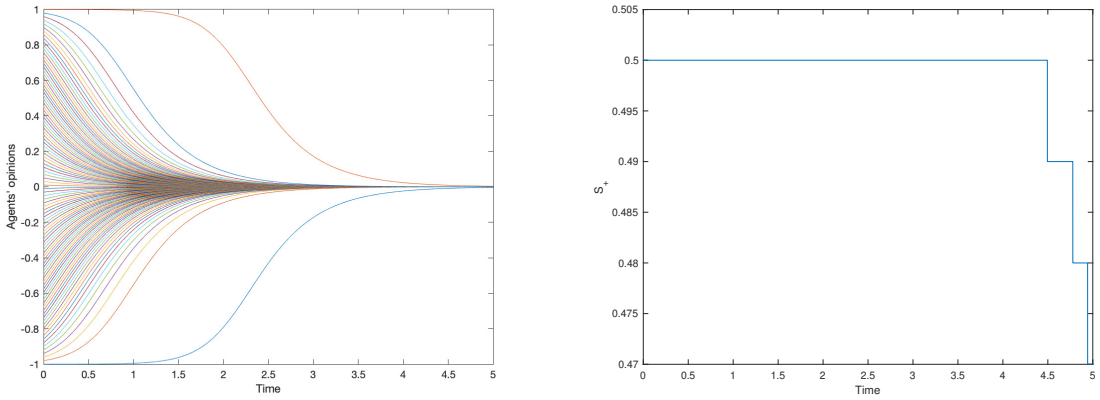

FIGURE 4. Time evolution of the opinions and of S_+ (Case 3).

4.2. **Strongly connected population.** In this second 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 5 (left) and the total number of non-zero entries is 2817. The initial number of posts of the agents of the population is the following: $b_1^0 = b_N^0 = 9$, $b_2^0 = b_{N-1}^0 = 8$, $b_3^0 = b_{N-2}^0 = 7$ and $b_i^0 = 1$ for all $i = 4, \dots, N-3$ (see Figure 5, right). The initial condition for the unknowns x_i is the same as in Cases 1, 2 and 3:

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

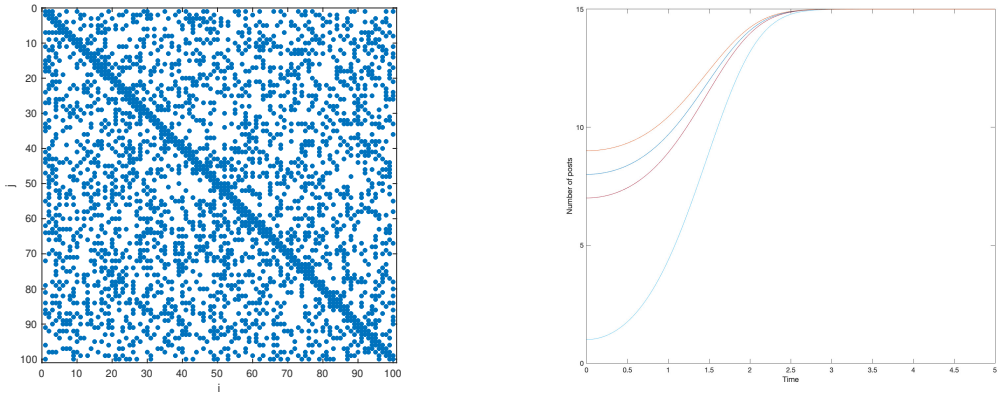

FIGURE 5. Interaction matrix and time evolution of the number of posts (Cases 4, 5 and 6).

4.2.1. *Case 4.* In Case 4, we suppose that the microblogging network's provider is neutral. We see that, even if the initial condition is the same as in Case 1, the geometry of the network representing the connexions between the individuals strongly modifies the behaviour of the system. In particular, the system tends, slower than in Case 1, to an equilibrium. This equilibrium is however different from zero (see Figure 6). 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 $S_+(5) = 0.12$.

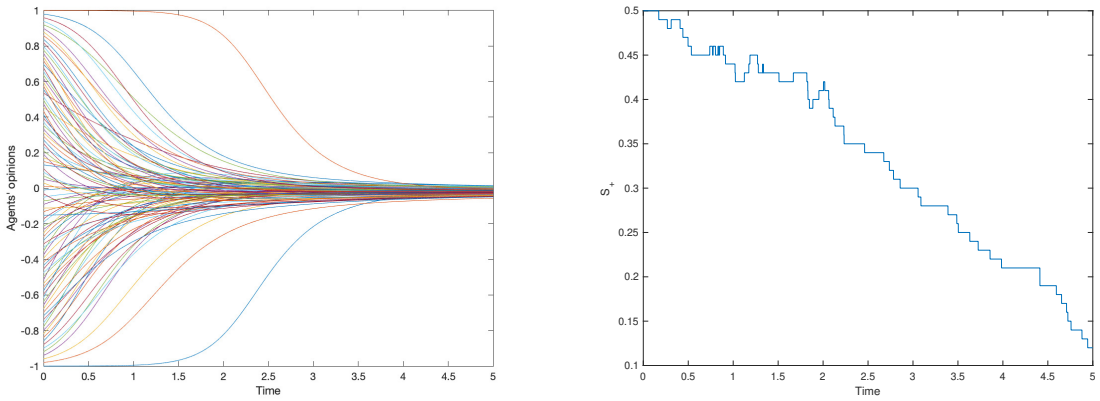

FIGURE 6. Time evolution of the opinions and of S_+ (Case 4).

4.2.2. *Case 5.* In this case, we study the effect of an hidden manipulation on the same population studied in Case 4. We suppose that the highlighting functions have the form

$$\psi_{i,j}(t, x_i(t), x_j(t)) = \begin{cases} 1 & \text{if } x_j(t) \geq x_i(t) \\ \dfrac{9}{10} & \text{otherwise.} \end{cases}$$

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 0.65, starting from the value $S_+(0) = 0.5$ (Figure 8) and with an interaction matrix which clearly favours the negative opinions, as shown in Case 4.

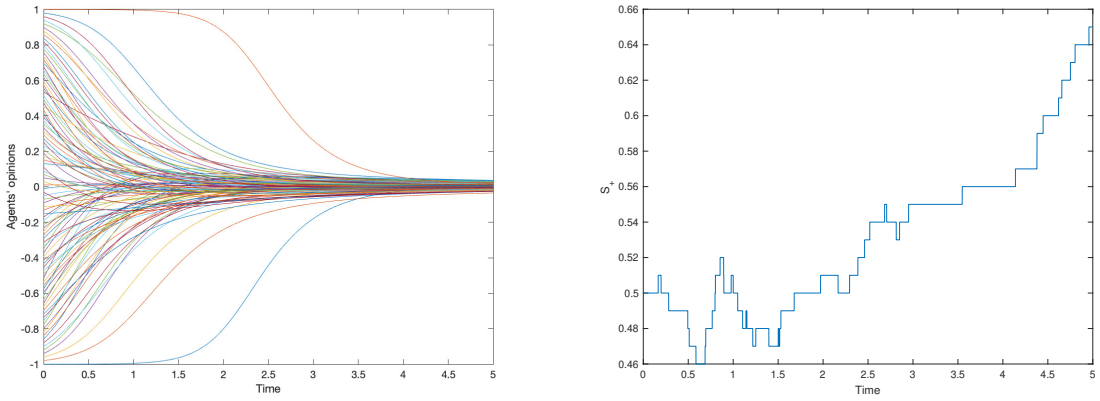

FIGURE 7. Time evolution of the opinions and of S_+ (Case 5).

4.2.3. *Case 6.* In this test, we slightly modify the form of the highlighting functions, in the sense that we turn off the manipulation effect in the simulation after two weeks. The precise form of the highlighting functions, which drive the system towards positive opinions, is the following

$$\psi_{i,j}(t, x_i(t), x_j(t)) = \begin{cases} 1 & \text{if } x_j(t) \geq x_i(t) \text{ or } t \geq 2 \\ \dfrac{9}{10} & \text{otherwise.} \end{cases}$$

We observe that a short-time manipulation (see [2] for a discussion about the timing of a perturbation on an opinion formation process) is enough for strongly modifying the collective behaviour of the population: we pass from the non-manipulated situation of Case 4, in which $S_+(5) = 0.12$, to a final result in which the population is favourable to the underlying question, with $S_+(5) = 0.56$.

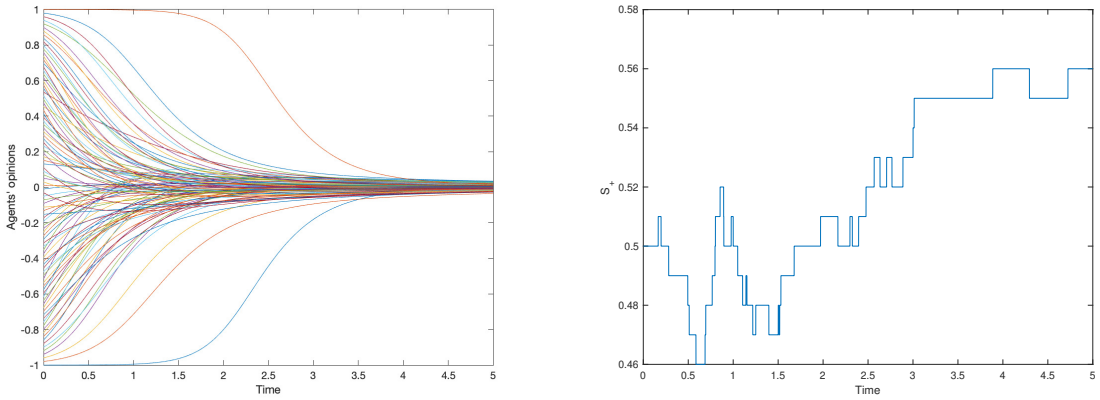

FIGURE 8. Time evolution of the opinions and of S_+ (Case 6).

4.3. **Clustered population.** The last series of tests 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 9. The total number of connexions of the network is equal to 849.

Initially, the population has average opinion equal to 0.3069. The fraction of the population with positive opinion at time $t = 0$ is $S_+(0) = 0.48$. 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 manipulation. In Case 8, the network highlights opinions which have negative sign whereas Case 9 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 manipulation effects, we suppose that all the agents have the same blogging activity (i.e. $b_i^0 = 1$ for all $i = 1, \ldots, N$).

4.3.1. *Case 7.* We first treat the situation without manipulation. In Figure 10 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 $S_+(0) = 0.48$ to $S_+(5) = 0.75$. The four detected clusters are consistent with the four interconnected sub-populations of the network.

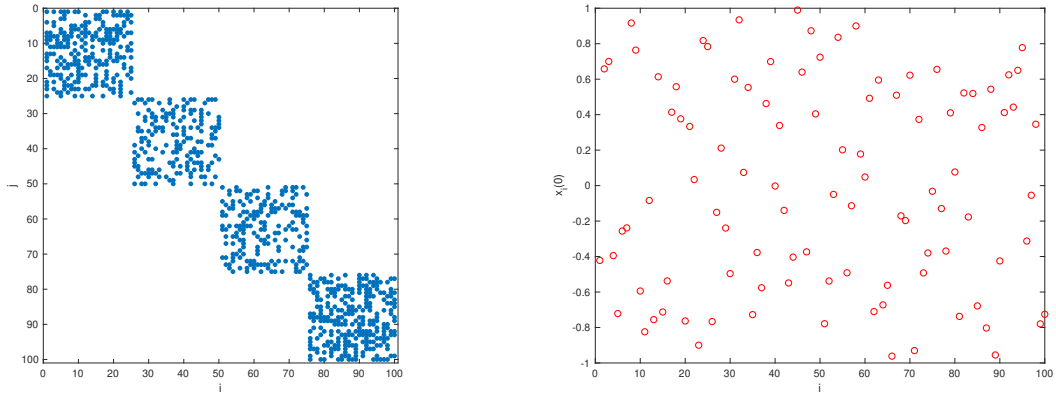

FIGURE 9. Interaction matrix (left) and initial condition (right), (Cases 7, 8 and 9).

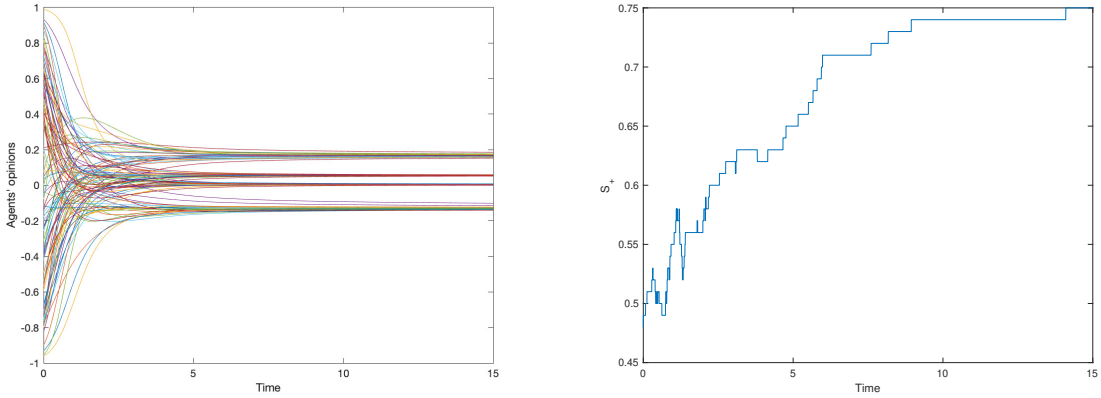

FIGURE 10. Time evolution of the opinions and of S_+ (Case 7).

4.3.2. *Case 8.* The manipulation effect is simulated by using highlighting functions of type

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

We underline that this form of the highlighting functions has a decisive effect in pushing the system towards negative opinions. In Figure 11, we note that the subpopulation S_+ decreases from $S_+(0) = 0.48$ to $S_+(5) = 0.25$. Moreover, the whole population reduces itself to four clusters, three of them are centred below zero. The result is very sensitive to the wights of the opinion in the highlighting functions. If we replace the value $4/5$ with $9/10$, the manipulation effect would be much weaker and the system would converge to a fifty-fifty equilibrium.

4.3.3. *Case 9.* The 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 12 shows that this strategy is less efficient than the strategy used in Case 8, 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

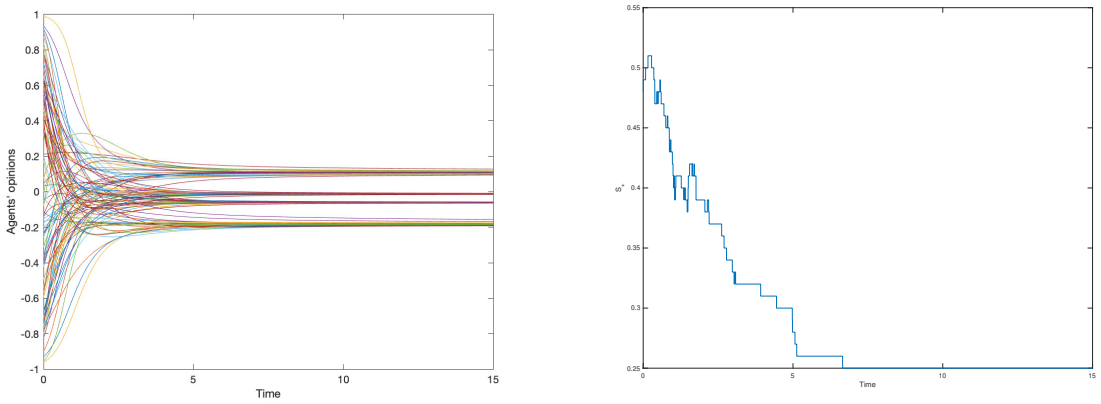

FIGURE 11. Time evolution of the opinions and of S_+ (Case 8).

for driving the system towards the equilibrium $S_+(5) = 0.5$, inducing in this way a fifty-fifty polarisation of the population, very far from the equilibrium, without manipulation, of Case 7.

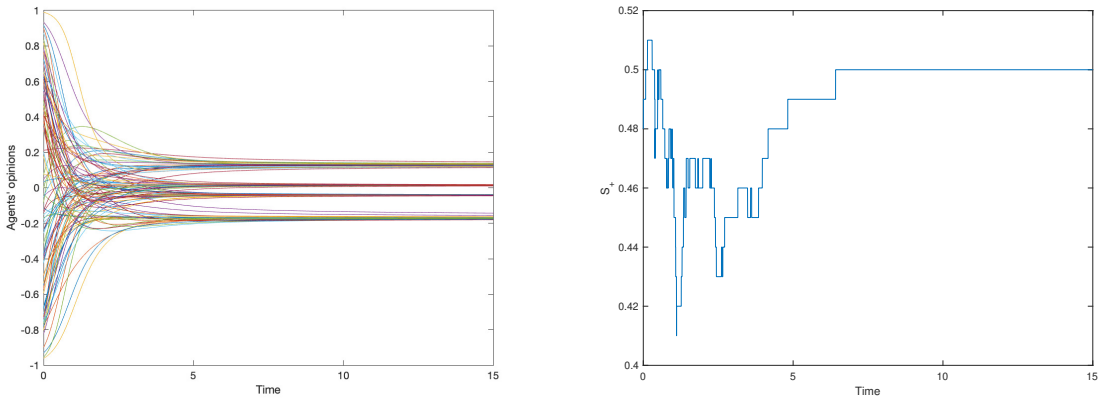

FIGURE 12. Time evolution of the opinions and of S_+ (Case 9).

5. Conclusion

We have studied some dynamics of opinion manipulation. 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 shown that hidden manipulation can have an important impact on the public opinion formation and that very mild interventions of the network owner can have major effects on the population.

A way for detecting such manipulation activities could be the design of an artificial isolated cluster of accounts, with initial conditions and blogging activity identical to those described in Case 1. If the behaviour of the system is not coherent with the dynamics of Case 1, it is possible that a hidden manipulation can be in progress. The fact that even temporary manipulation can be detected, as in Case 3, shows the sensitivity of this strategy for discovering manipulation attempts.

Acknowledgements. This work has been carried out in the framework of the projects *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.

References

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

[2] 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.

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

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

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

[6] 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.

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

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

[9] 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.

[10] Talcott Parsons, Edward A Shils, et al. Values, motives, and systems of action. *Toward a general theory of action*, 33:247–275, 1951.

[11] 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.

[12] Thomas J Scheff. Toward a sociological model of consensus. *American Sociological Review*, pages 32–46, 1967.

(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