THE DARK SIDE OF GOSSIPS: HINTS FROM A SIMPLE OPINION DYNAMICS MODEL

Guillaume Deffuant, Ilaria Bertazzi, Sylvie Huet

Advances in Complex Systems

The dark side of gossips: Hints from a simple opinion dynamics model

Guillaume Deffuant, Ilaria Bertazzi, Sylvie Huet
LISC, Irstea, France

Received (received date)
Revised (revised date)

We consider a simple model of agents modifying their opinion about themselves and about the others during random pair interactions. Two unexpected patterns emerge: (1) without gossips, starting from zero, agents opinions tend to grow and stabilize on average at a positive value; (2) when introducing gossips, this pattern is inverted; the opinions tend to decrease and stabilize on average at a negative value. We show that these patterns can be explained by the relative influence of a positive bias on self-opinions and of a negative bias on opinions about others. Without gossips, the positive bias on self-opinions dominates, leading to a positive average opinion. Gossips increase the negative bias about others, which can dominate the positive bias on self-opinions, leading to a negative average opinion.

Keywords: Opinion dynamics; gossips; self-opinion; positivity bias.

1. Introduction

Gossips are the subject of a large body of research from different disciplines (see a review in [17]) which generally emphasize their social utility. Indeed, gossips help reputation management [11, 27] and to solve social problems such as propagating information about cheaters or about potential partners or punishing deviations from the social norm. Also, gossips can introduce indirect altruistic behaviours because agents are motivated to maintain a good reputation [26]. Nevertheless, a few studies [4] observe negative effects of gossips on group cohesion.

This paper shows this dark side of gossips in a simple model of opinion dynamics and suggests that it is deeply rooted in social interactions. As far as we know, this is the first time that this issue is addressed through a modelling approach.

The model comprises a set of agents, each one holding an opinion (a real number between -1 and +1) about herself and about each other agent. During random dyadic encounters, each agent modifies her opinions about both agents in the couple, under the influence of the other. The influence is attractive; each agent tends to get her opinion closer to her evaluation of the one of her interlocutor, with three specific additional assumptions:

- The evaluation of others' opinion is noisy, accounting for random communication mistakes. The average noise is zero, expressing that the mistakes

are not biased.
- Agents are more strongly attracted by the opinions of agents that they value higher than themselves and less attracted by the opinions of agents that they value less than themselves. Overall the influence function has a sigmoïd shape.
- When lower than -1, opinions are truncated to -1 and when higher than +1, they are truncated to +1.

Gossips can be added into the model, in which case, at each encounter agents influence each others' opinions not only about themselves, but also about other agents (chosen at random). This model is a simplified version of the Leviathan model [9, 21], without the process called vanity[a]. Note that the vanity process can account for the negative reactions generated by agents who are felt to have a too high opinion of themselves, if these agents also under-evaluate others. This type of reaction is absent from the simplified model considered in this paper: an agent having a high value of herself tends to convince the others of this high value. Indeed, our main objective is to better understand the phenomena, already complex, taking place in the simplified model.

This model can be related to the large family of opinion dynamics models [14, 16, 25, 5, 7, 18, 13] (for a recent review see: [12]). Indeed, the interactions on opinions follow the classical principle of attractive influence often found in these models. However, this model shows the rather uncommon feature that the agents' opinions are about each other, whereas most of the opinion dynamics models assume agents talking about things and objects, for instance about products or movies (despite a few exceptions such as [1, 3]).

Yet, it can be argued that opinions about people are more important and more interesting than opinions about things. First, it has been observed that most casual conversations are about the speakers themselves or about other people they know [10, 15, 31]. Then, it is a common observation that, when speaking about things, people often seek to send messages about themselves or about others. Finally, the opinion dynamics about people generate spontaneous hierarchical social structures and define agents' views of themselves within them. Therefore a better understanding of these processes could shed a new light on deep psychological and social dynamics.

Two categories of models about gossips and reputations are proposed in [30]: the cognitive models which focus on the dynamics of mental states of the agents, and the game theoretical models in which the behaviour of the agents in the outside world is decisive for their reputation. Our model belongs to the first category: the

[a]The principle of the vanity process is to be grateful when you feel highly valued and angry when you feel lowly valued: agent 1 increases her opinion about agent 2 when agent 1 thinks that agent 2 has an opinion about agent 1 which is higher that agent 1's self-opinion and agent 1 decreases her opinion about agent 2 when agent 1 thinks that agent 2 has an opinion which is lower that agent 1's self opinion.

agents discuss only about their respective opinions, without any reference to their behaviour.

This corresponds to the willingness to simplify the model as much as possible, in order to identify clear effects [6, 8]. The model thus does not claim to represent realistic situations. In particular, it assumes that all agents are initially interchangeable. This can be seen as a neutral hypothesis, like the one proposed in ecology by Hubbell [20]. The differentiation between agents comes thus only from the history of random interactions. Moreover, in the most simplified version of the model (without gossips), we suppose that the agents talk only about themselves and about their interlocutor, which of course is rarely the case.

However, we claim that the core assumptions of the model on the dynamics of opinions are rooted in well established psycho-sociological knowledge. Indeed, the tendency to follow each others' attitudes is attested by numerous experiments and the main basis of most opinion models (with many nuances and sophistication, of course). The tendency to give more influence to valued people is a component of one of the best confirmed theories in social psychology [28]. Our goal therefore, is to identify the general effects of these mechanisms in idealized situations where their role should be more easily understood. In sum, our approach is to study in depth the emergence of pure patterns in oversimplified cases, with the hope that this will help understand more complicated settings.

We focus on two patterns:

- the *positive drift of the model without gossips*: starting from all opinions at zero, the average opinion tends to grow for a while and then fluctuates around a stable positive value. Each agent tends to have a positive opinion about all the others, including herself.
- the *negative drift of the model with gossips*: when gossips are added to the model, on the contrary, starting from all opinions at zero, the average opinion tends to decrease and then fluctuates around a stable negative value. Each agent tends to have a negative opinion about all the others, including herself.

These patterns are rather surprising because at a first glance, the equations of the model appear to be symmetric and the noise is unbiased, therefore the distribution of opinions is expected to be balanced between the negative and positive sides.

The main contribution of this paper is to propose an explanation of these patterns which relates to two statistical biases appearing in the model:

- a positive bias on self-opinions: when agent 1 keeps a constant opinion about agent 2, agent 2's self-opinion is on average a bit higher than agent 1's opinion about her;
- a negative bias on opinions about others: when agent 1 keeps a constant self-opinion, the opinion of agent 2 about agent 1 is on average a bit lower than agent 1's self-opinion.

When there are no gossips, the positive bias on self-opinions dominates and tends to drive the opinions upwards. Gossips increase the negative bias, sometimes sufficiently to overcome the positive bias.

The following section firstly describes the model and the patterns under scrutiny; then, we observe in details simplified simulation experiments with two and then with more agents that suggest how the effects of the positive and negative biases may lead to the positive and negative drifts. The paper concludes with a discussion on the possible applications of the model to empirical cases.

2. The Model

The model includes n agents, each agent M (Me) having an opinion a_{MY} about each agent Y (You) including herself; the opinions are real values between -1, the worst opinion, and +1, the best opinion. Initially, all opinions are set to 0: agents have a neutral opinion about all the others at the beginning of the simulation.

Graphically, we represent agents' opinions as a matrix, in which row M is the array of opinion that agent M has on the other agents Y and, column M is the opinions all agents Y have about M. The right-bottom top-left diagonal shows the agent self-opinions. Positive opinions are represented with red shades and negative opinions with blue shades. Lighter shades are used for feeble opinions (close to 0), and they get darker as the opinion becomes more polarized towards -1 or +1, as represented in Figure 1. We define the *reputation* r_M of agent M as the average of the others' opinions about her.

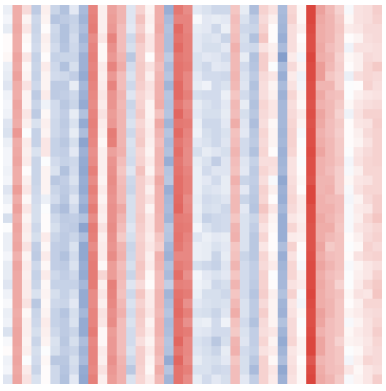


Fig. 1. Opinion matrix. Each row represents an agent's opinion over the population, with the right-bottom to top-left diagonal being self-opinions.

2.1. *Dynamics without gossips*

At each time step two randomly chosen agents M and Y have an encounter and influence one another. The change Δa_{MY} of opinion of M about Y leads a_{MY} to get closer to a noisy evaluation of Y's self-opinion, according to the following equation, in which $R(\delta)$ designates a uniformly drawn number between $-\delta$ and δ:

$$\Delta a_{MY}(t) = p_{MY}(t)(a_{YY}(t) - a_{MY}(t) + R(\delta)). \tag{1}$$

In this equation, the change of a_{MY} is influenced by a_{YY} (Y's self-opinion) and the amplitude of this influence is modulated by a propagation coefficient p_{MY}, that will be explained shortly. The random variable $R(\delta)$ expresses the mistakes done by M when evaluating Y's self-opinion.

At the same time, the change Δa_{MM} of M's self-opinion tends to get a_{MM} closer to a noisy evaluation of Y's opinion about M, in a similar way:

$$\Delta a_{MM}(t) = p_{MY}(t)(a_{YM}(t) - a_{MM}(t) + R(\delta)). \tag{2}$$

The changes of opinions of Y on M obey the same rules, with a different propagation coefficient p_{YM} (Notice that $M \neq Y$; an agent cannot be paired with herself).

The *Propagation Function* p_{MY} represents how much M is influenced by Y. This coefficient is a sigmoïd function of the difference between M's opinion about Y (a_{MY}) and M's self-opinion (a_{MM}). This function tends to 1 if M values Y much higher than herself, and tends to 0 when M values Y much lower than herself:

$$p_{MY}(t) = \frac{1}{1 + \exp\left(\frac{a_{MM}(t) - a_{MY}(t)}{\sigma}\right)}. \tag{3}$$

Rephrasing, the function p_{MY} expresses the hypothesis that the more M perceives Y as superior to herself, the more Y is influential on M. This assumption is grounded in social psychology literature, and the sigmoïd function is classically used as a smooth threshold [22]. Figure 2 represents the graph of p_{MY} for different values of σ; notice that p_{MY} is always between 0 and 1.

The model therefore relies on two parameters, each of them spanning continuously from 0 to 1:

- σ defines the shape of the propagation function p_{MY}; if σ is very small, the function is very tilted, meaning that agents are subject to high influence from the ones who they evaluate better than themselves and they almost completely disregard the opinions of the ones considered lower.
- δ represents the amplitude of the uniformly distributed errors that perturb the evaluation of others' opinions. This noise expresses that an agent M cannot directly access the opinion of another, Y, and may often make errors in this evalu-

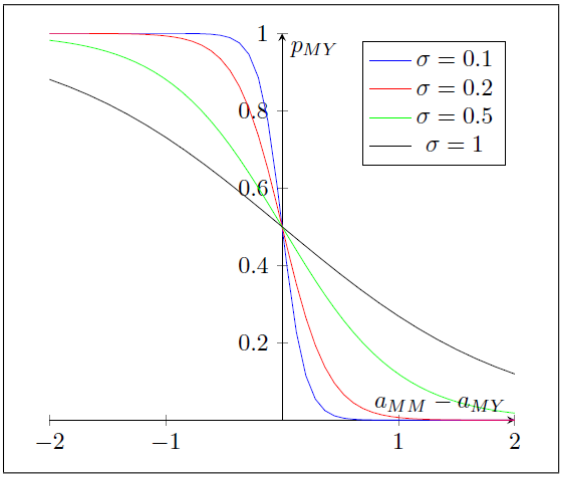


Fig. 2. Propagation Function p_{MY} with different values of σ. The influence given by M to Y decreases when M's self-opinion increases.

ation. In other words, the noise accounts for imperfect information transmission. Note that it is the main engine of the dynamics. Without it, from an initialization of all opinions at zero, there would be no opinion change at all.

In this paper, we limit our study to the model with synchronous update: at each encounter all the changes of opinions are first computed and then the opinions are modified simultaneously:

$$a_{MY}(t+1) = a_{MY}(t) + \Delta a_{MY}(t) \tag{4}$$

$$M, Y \in \{1, ..., N\} \tag{5}$$

The asynchronous model shows broadly the same emerging patterns but the positive or negative drifts are less pronounced.

2.2. *Introducing Gossips*

Introducing gossips in the model means that, when two agents meet, they do not only talk about their mutual opinions; they also exchange opinions over k other agents. The influence on opinion about others follows the same equation as before. Let H (Her) be the agent being object of gossip, M gets influenced by Y's opinion on H:

$$\Delta a_{MH}(t) = p_{MY}(t)(a_{YH}(t) - a_{MH}(t) + R(\delta))\,. \tag{6}$$

The model with gossips includes one more parameter, k, which represents the number of other agents that agent Y talks about when she discusses with M. These k

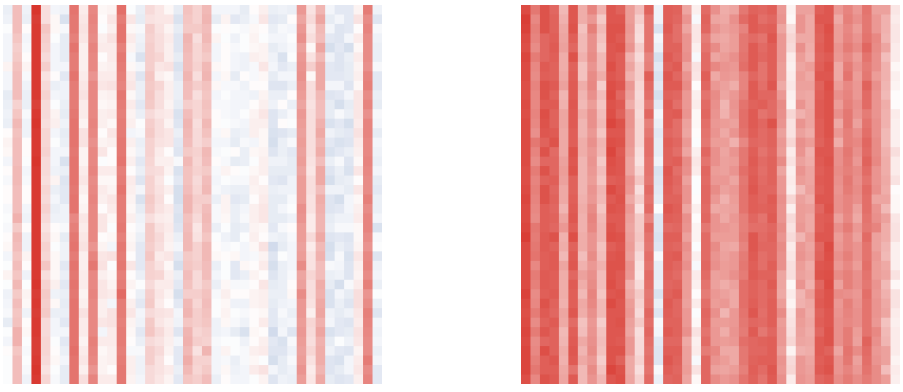
agents are chosen at random. Again, in this case, at each encounter all the discussed opinions are modified simultaneously.

2.3. *Emerging patterns: positive drift undermined by gossips*

An extensive exploration of the trajectories of the simplified model without gossips for different parameter values is reported in [21]. We concentrate our current analysis in a part of the parameter space where parameter σ is such that the propagation function has an intermediate slope (not too smooth and not too sharp), a typical value is $\sigma = 0.3$ and the noise parameter δ is of the same order as σ (we choose $\delta = 0.2$). For such parameter values, the simulations exhibit two main features:

- The positive drift of the model without gossips: starting all from 0, the opinions grow and then stabilize on average at a significantly positive value;
- The opinions tend to be similar in the columns of the matrix: a_{MH} and a_{YH}, with $M \neq Y$, tend to be close. This feature has already been observed in the Leviathan model; it is related to the well-known convergence to the mean value of attractive opinion dynamics which takes place here because the sigmoid function is never very close to 0 (see [9] for details).

Figure 3 shows two opinion matrices at different time steps for $n = 40$ agents, $\delta = 0.2$ and $\sigma = 0.3$, for a simulation of the model without gossips. The two features appear: the shades generally tend to red (the average opinion is significantly positive), and the columns tend to have homogeneous colours.



(a) (b)

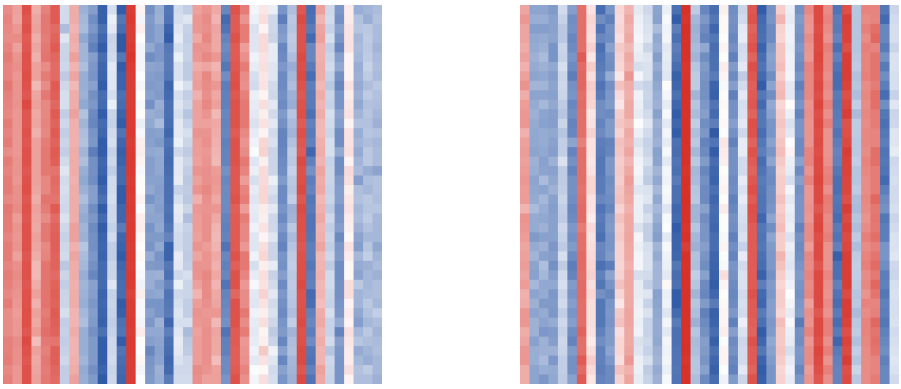
Fig. 3. Simulation run without Gossips, after 100,000 (a) and after 1,000,000 (b) time steps. $\delta = 0.2$ and $\sigma = 0.3$.

For the same parameter values, introducing gossips undermines the positive drift and, in the long run leads to a negative drift instead. In this case, as shown in Figure 4, alongside with lighter shades of red (positive reputations) there are also darker

shades of blue, which represent negative reputations for the respective agents. The average of the opinions is negative for the model with gossips whereas it is positive for the model without.

When adding gossips in a situation where the opinion is positive on average because of the positive drift without gossips, the average opinion decreases, becomes negative and then fluctuates around a negative value. If the gossips are stopped at this moment, the positive drift starts again and the average opinion comes back to the initial positive value.

These observations are valid for any value of $k > 1$. However, the number of agents should be higher than 7. Below this number, there is no positive drift without gossips.



(a) (b)

Fig. 4. Simulation run with Gossips, after 100,000 (a) and after 1,000,000 (b) time steps. $\delta = 0.2$, $\sigma = 0.3$ and $k = 5$.

3. Explaining the positive drift

As already noticed in the Leviathan, agents tend to have a self-opinion (a_{MM}) that is higher than the one the others have on them (a_{YM}). In this section we will argue that this statistical regularity can be disentangled into two phenomena, a positive bias over self-opinion and a negative bias on the opinion over the others. The joint effect of these two biases is at the core of the positive and negative drifts.

3.1. *Positive bias on self-opinion and negative bias on opinion about others*

We focus firstly on a simplified case with two agents only, M and Y. The positive bias on self-opinion can be put in evidence when supposing that only the self-opinion a_{MM} is changing over time while all the other opinions a_{YY}, a_{MY}, a_{YM} are fixed,

with $a_{MY} = a_{YY}$. When a_{MM} fluctuates above a_{YM}, p_{MY} decreases, and M gives less influence to Y; the reverse happens in the opposite case. Therefore M's opinions tend to be more stable after positive fluctuations, because M is less prone to listen to Y, and less stable after negative fluctuations for opposite reasons. Overall the average opinion $< a_{MM} >$ over a large number of interactions is therefore slightly higher than the opinion a_{YM} of Y about her.

The negative bias on the opinion about others designates the opposite tendency when only a_{YM} changes over time, while all the other opinions are fixed. Indeed, in this case, when a_{YM} fluctuates below the fixed value of a_{MM}, the influence of M decreases and thus a_{YM} becomes more stable. Therefore, the average opinion $< a_{YM} >$ over a large number of interactions tends to be lower than the fixed value of a_{MM}.

Figure 5 shows these biases for $a_{YY} = 0.0$ (They are simply shifted for different values of a_{YY}). The positive bias on self-opinion is computed for 100 fixed values of a_{YM} regularly distributed from -0.8 to $+0.8$. The graph shows $< a_{MM} > -a_{YM}$, where $< a_{MM} >$ is the average value of a_{MM} over $500,000$ interactions. Similarly for the negative bias on the opinion about others, we plot $< a_{YM} > -a_{MM}$, with a_{YM} only varying for 100 fixed values of a_{MM} regularly distributed from -0.8 to $+0.8$ (and also 500000 interactions at each fixed value).

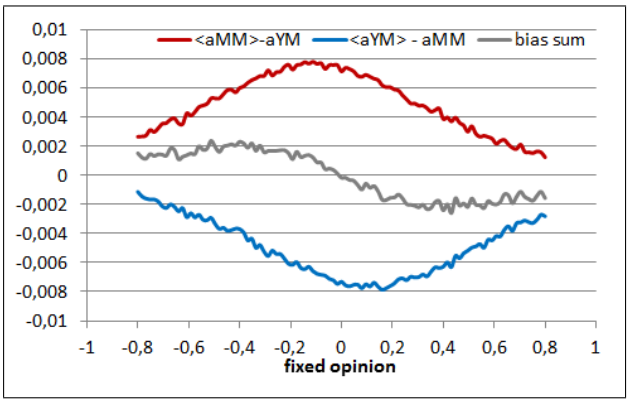


Fig. 5. Positive and negative biases for $a_{YY} = 0.0$ computed on 500000 interactions. The grey curve is the sum of both biases. It cuts the horizontal axis at the value of a_{YY}

Figure 5 shows also the sum of the biases. This sum is made supposing that the values of the x-axes are approximately the same for the positive and negative biases, in other words that the biases are small. Making this approximation, it appears that the sum of biases is positive below the value of a_{YY} and negative above. Indeed the strength of the biases increases with the value and the slope of the propagation function which explains the asymmetric bell shape of the graphs. For very negative values of a_{YM}, the propagation function is high (close to 1) but its slope is very

small, thus the positive bias is not very high, then when increasing a_{YM}, the slope of the propagation function increases while its value decreases, and the positive bias reaches its maximum for a_{YM} a bit below the value of a_{YY}, then it decreases because both the slope and the value of the propagation function decrease. For the negative bias, the analysis is similar, except that the bias reaches its maximum absolute value for a_{MM} slightly above the value of a_{YY}, and then this absolute value decreases to a value which is smaller than the one of the positive bias.

The graph of the sum of the biases shows that the positive bias dominates when a_{YM} (supposed almost equal to a_{MM}) is below the value of a_{YY}, while the negative bias dominates when a_{YM} is above this value.

Note that these conclusions hold only for the chosen set of parameter values (σ such that the slope of the sigmoïd function is intermediate and noise of the order of σ). Indeed, when the slope of the propagation function is lower (larger σ), the difference between the positive and the negative biases is smaller, hence they tend to neutralise each other and the positive drift is weaker. When the slope of the propagation function is too small, then the higher agents are no more influenced by the lower ones and other patterns take place, see [21]). Similarly, if the noise is too high, the hypothesis of relative stability of the positions of the agents with respect to each other is no longer valid.

3.2. *The higher influence of the weaker agent in the long term*

We are now supposing that a_{MM}, the self-opinion of M, and a_{YM}, the opinion of Y about M are varying together using the model equations, while $a_{MY} = a_{YY}$ are fixed[b].

In order to better capture robust trends in the evolution of opinions, we average them in a moving window of w time-steps. For instance, considering a_{YM}, we define its moving average $\bar{a}_{YM}(t)$ as follows:

$$\bar{a}_{YM}(t) = \frac{\sum_{\tau=-w}^{w} a_{YM}(t+\tau)}{2w+1}. \tag{7}$$

Moreover, we compute time averaged opinion changes after v time steps in order to identify more easily robust tendencies. We thus define $\Delta\bar{a}_{YM}(t)$ as:

$$\Delta\bar{a}_{YM}(t) = \bar{a}_{YM}(t+v) - \bar{a}_{YM}(t). \tag{8}$$

Similarly, we compute the time window averaged values $\bar{a}_{MM}(t)$ and their changes $\Delta\bar{a}_{MM}(t)$.

We perform a large number of interactions and we compute the distributions of the different variables as a function of $\bar{a}_{YM}$. In the reported experiment below, we

[b]The results are not significantly different when a_{MY} is also modified according to the model equations (not fixed). Indeed, a_{MY} is then fluctuating around a_{YY} and the impact on the dynamics of a_{MM} and a_{YM} compared with fixing $a_{MY} = a_{YY}$ is negligible.

choose $w = 60$ and $v = 30$ because with these values the distributions of $\Delta\bar{a}_{MM}$ and $\Delta\bar{a}_{YM}$ are close to each other, in accordance with the observed convergence of the variations in columns. It suggests that these values grasp the main tendencies of the time evolution. We keep these values of w and v for all the other experiments.

Figure 6, shows $< \Delta\bar{a}_{YM}(t) >$ the average value of $\Delta\bar{a}_{YM}(t)$ for values of $\bar{a}_{YM}$ located in one of $h-1$ regularly distributed intervals on the $[-1, +1]$ axis ($h = 100$ in the graphs), when doing a large number of interactions (50 million). More precisely, for each interval $I_i = [-1 + \frac{2i}{h}, -1 + \frac{2i+2}{h}]$, with $i \in \{1, ..., h-1\}$, during the simulation, each time $\bar{a}_{YM}(t) \in I_i$, we store the corresponding value of $\Delta\bar{a}_{YM}(t)$. At the end of the simulation, for each interval I_i, we compute $< \Delta\bar{a}_{YM}(t) >_i$, the average of the stored values of $\Delta\bar{a}_{YM}(t)$ for this interval and Figure 6 shows the graphs of $< \Delta\bar{a}_{YM}(t) >_i$, $i \in \{1, ..., h-1\}$.

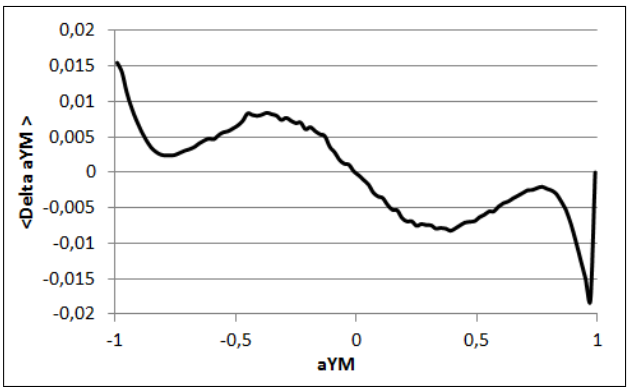


Fig. 6. $< \Delta\bar{a}_{YM}(t) >_i$ as a function of $\bar{a}_{YM}(t)$ computed on 50 Million interactions between two agents M and Y, $a_{YY} = a_{MY} = 0.0$ fixed, and 100 intervals on the $\bar{a}_{YM}$ axis.

Figure 6 shows that $< \Delta\bar{a}_{YM}(t) >$ is positive for $\bar{a}_{YM}(t)$ below a_{YY} and negative for $\bar{a}_{YM}(t)$ above a_{YY}. Therefore, when $\bar{a}_{YM}$ is below a_{YY}, the value of $\bar{a}_{YM}$ tends to increase on average, whereas it is the opposite when $\bar{a}_{YM}$ is above a_{YY}.

This result can be explained with the biases shown on Figure 5. When $\bar{a}_{YM}$ is lower than a_{YY}, we have seen that the positive bias on a_{MM} is larger than the negative bias on a_{YM}. In other words, M keeps repeating to Y "I'm better than what you think", more strongly than Y keeps repeating to M "You are worse than what you think". Y tends to get convinced progressively which leads to increase (on average) both a_{MM} and a_{YM}. Similarly, when $\bar{a}_{YM}$ is higher than a_{YY}, the negative bias on a_{YM} is larger than the positive bias on a_{MM} and things are inverted, leading to the average downward tendency of $\bar{a}_{YM}$. The differences of amplitudes of the biases are due to the shape of the propagation function as analysed in section 3.1.

Therefore, in this analysis, the positive $< \Delta\bar{a}_{YM}(t) >$ for $\bar{a}_{YM}$ lower than a_{YY} and the negative $< \Delta\bar{a}_{YM}(t) >$ for $\bar{a}_{YM}$ higher than a_{YY} shown on figure 6 are due to the larger bias of the agent with the lower self-opinion. Indeed, the positive bias

on a_{MM} is larger when $\bar{a}_{YM}$ and $\bar{a}_{MM}$ (supposed closed to each other) are lower than a_{YY} and the negative bias on a_{YM} is larger for high values of $\bar{a}_{YM}$ therefore when a_{MM} is higher than $\bar{a}_{YY}$. Therefore, the dominated agent (with the lower self-opinion) is less influential in the short term, but in the long term, her steadily stronger bias actually tends to drive the evolution of a_{YM}.

3.3. *Collective neutralizing of the negative biases*

We observed that the positive drift takes place only when there are at least 7 agents and the previous experiment confirms that the positive drift does not take place with two agents. We are now considering an agent M interacting with $n-1$ agents Y, each with a fixed self-opinion a_{YY} and we run similar experiments as the ones we performed with a couple of agents. The reputation $r_M(t)$ of agent M is the average of the opinions that the others have on her:

$$r_M(t) = \frac{\sum_Y a_{YM}}{n-1}. \tag{9}$$

Again, we consider the average of the reputation in a time window of size $w = 60$, defined as previously:

$$\bar{r}_M(t) = \frac{\sum_{\tau=-w}^{w} r_M(t+\tau)}{2w+1}. \tag{10}$$

We first consider the simple case of two values of $a_{YY} \in \{-0.5, 0.9\}$. Figure 7 shows $< \Delta\bar{r}_M(t) >_i$ the average variation of $\Delta\bar{r}_M(t) = \bar{r}_M(t+v) - \bar{r}_M(t)$ ($v = 30$), the average values $< \Delta\bar{a}_{YM}(t) >$ of the variations of the two $\bar{a}_{YM}$ for values of $\bar{r}_M(t)$ located in the intervals I_i previously defined over a simulation of 50 million time steps. It appears on this figure that $< \Delta\bar{r}_M >_i$ is positive for most intervals I_i, except when $\bar{r}_M$ is higher than 0.7.

Figure 8 shows the average variations of the $< \Delta\bar{a}_{YM} >_i$, for the two agents Y (note that the average of these two values is the average variation of the reputation). It appears that the variation of $\bar{a}_{YM}$ from the agent such that $a_{YY} = -0.5$ is always positive except for values of $\bar{r}_M$ higher than 0.7, with an important contribution for values of $\bar{r}_M$ around -0.5. The contribution to the variation of $\bar{r}_M$ from the agent such that $a_{YY} = 0.9$ is very slightly negative around $\bar{r}_M = -0.5$ and constantly increasing until a strong decrease close to $\bar{r}_M$ equals 0.7. The shape of these curves are significantly different from the one found when only two agents interact, which suggests that the combined effect of several agents is not the simple cumulation of the effect of pair interactions. In particular, the negative tendency found after the value of a_{YY} in pair interactions is not visible for the curve $< \Delta\bar{a}_{YM} >$ for $a_{YY} = -0.5$. This is presumably an effect of the agent with a higher a_{YY}, which neutralises this negative tendency.

As a result, the reputation tends to grow on average, except when it reaches the values above 0.7 where the border effect due to the truncation of the opinion at +1 is

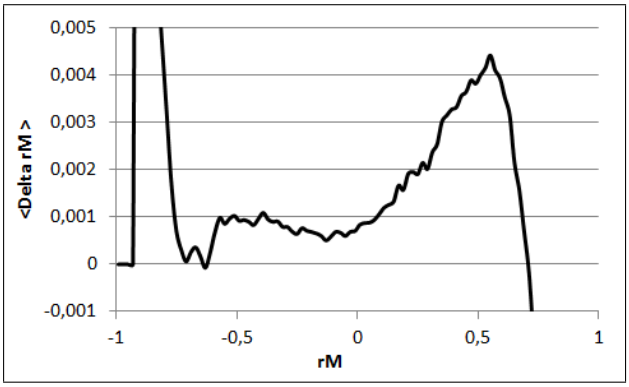


Fig. 7. Average variation of the time window averaged reputation $< \Delta\bar{r}_M >_i$ for 50 million iterations and 100 intervals on the $\bar{r}_M$ axis. Two You agents, $a_{YY} \in \{-0.5, 0.9\}$.

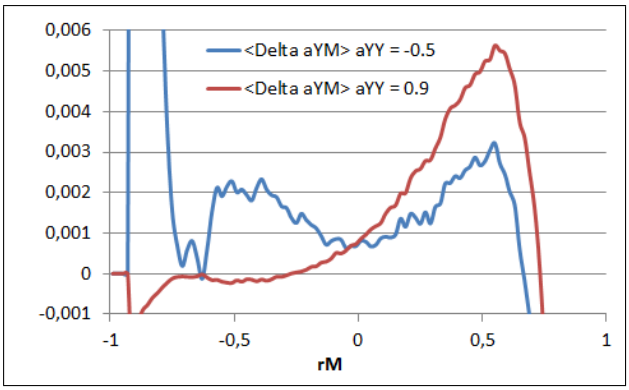


Fig. 8. Average variation of time window averaged opinions $< \Delta\bar{a}_{YM} >_i$, for 50 million iterations and 100 intervals of the $\bar{r}_M$ axis, for $a_{YY} = -0.5$ and $a_{YY} = 0.9$.

probably dominant. Overall, when neglecting this border effect, the distribution of $r_M(t)$ shows a positive drift. Nevertheless, it is important to keep in mind that the values $< \Delta\bar{r}_M >_i$ are averages over a large number of trajectories crossing interval i and that many of them are downward despite the average upward tendency.

However, generally with only two Y agents with fixed a_{YY}, the positive drift on r_M does not take place. For instance, if instead of $a_{YY} \in \{-0.5, 0.9\}$ we take $a_{YY} \in \{-0.2, 0.5\}$, the graph of $< \Delta\bar{r}_M >$ is not so positive for negative values of $\bar{r}_M(t)$ and if the second value is lower, then the negative bias appears for lower values of $\bar{r}_M(t)$. Therefore, the positive drift requires that the distribution of a_{YY} keeps covering the axis from -0.5 to 0.9. For instance, the evolution of $\Delta\bar{r}_M$ shown on figure 9 for 8 agents y with fixed a_{YY}, distributed uniformly from -0.6 to 0.8, generates a positive drift.

When the number of agents is large enough, the fluctuations of opinions (because

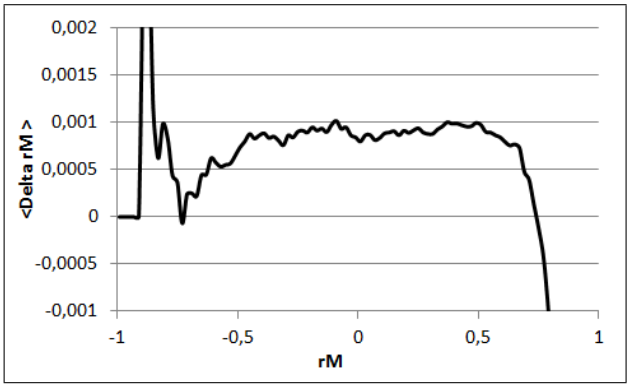


Fig. 9. Average variation of the time window averaged reputation $< \Delta\bar{r}_M >_i$ for 50 million iterations and 100 intervals. $a_{YY} \in \{-0.6, -0.4, -0.2, 0.0, 0.2, 0.4, 0.6, 0.8\}$.

of the noise) lead the agents' self-opinions to cover of the axis of opinions sufficiently to ensure a generally positive function $< \Delta\bar{r}_M >$ and a positive drift. When the number of agents is too small (experimentally, smaller than 7) the fluctuations in the agents' self-opinions do not ensure adequate covering of the opinion axis and the positive drift does not take place.

Finally, our analysis suggests that the positive drift is generated by the combination of the positive biases on a_{MM} which together overcome the negative biases on a_{MY} when the distribution of agents self-opinions covers the opinion axis adequately. Indeed, the positive bias generated from the agents of higher self-opinions a_{YY} seems to neutralize the negative bias generated by the lower a_{YY}. We can therefore again interpret the process as being driven by the weak agents (with low reputations) who collaborate in convincing the ones with high reputations to increase their opinions.

4. Explaining the negative drift with gossips

4.1. *Gossips as additional noise on interactions about others*

As stressed out in Section 2, introducing gossips undermines the positive drift and can even revert it to a negative drift. We change the experiment described in Section 3.3, by adding interactions between two Y agents talking about M (with 8 fixed agents and one gossip happening each time step). The average reputation change $< \Delta\bar{r}_M >_i$ is presented in figure 10 with the reputation change without gossips (same as in Figure 9). It appears that, with gossips, $< \Delta\bar{r}_M >_i$ becomes negative, indicating a downward tendency of the average reputation.

This effect is easier to understand in the light of the analysis of the positive drift; indeed the gossips tend to increase strongly the fluctuations of a_{YM} because these gossips also come from influential agents and strongly modify a_{YM}. The gossips do not increase the fluctuations of a_{MM}, because the interactions about the self-

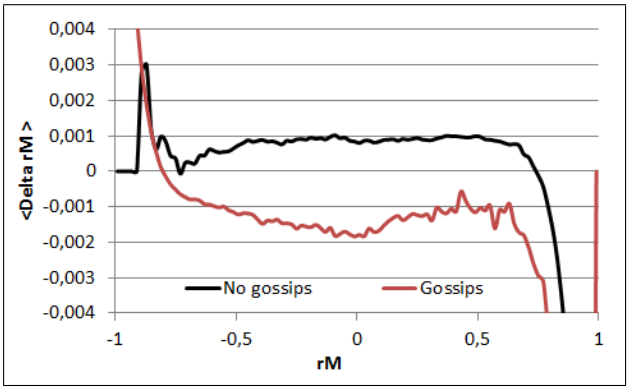


Fig. 10. Average variation of the window averaged reputation $< \Delta \bar{r}_M >_i$ for 30 million iterations and 100 intervals with gossips. $a_{YY} \in \{-0.6, -0.4, -0.2, 0.0, 0.2, 0.4, 0.6, 0.8\}$.

opinion are not directly modified by the gossips. Therefore, the gossips strengthen the negative bias without modifying the positive bias.

In this view, the gossips appear simply as an additional source of noise on the opinion about others. We run another version of the standard, simplified model, where at each time step, each agent is subject to an additional, independent noise over the opinion she has over one random other agent. This noise is always the same in amplitude but with a random sign, in this case ± 0.04. Figure 11 shows the average reputation change in the time window for the model with the additional noise of ± 0.04 compared with the model without this noise and with the model with gossips, in the case of an agent M interacting with 8 agents Y, with $a_{YY} \in \{-0.6, -0.4, -0.2, 0.0, 0.2, 0.4, 0.6, 0.8\}$. The tendency of the reputation is similarly downward ($< \Delta \bar{r}_M >_i$ negative) for most of the values of $\bar{r}_M$ for the models with additional noise or with gossips, while it is upward for the model without noise and without gossips. Again this can be interpreted as the effect of the negative bias on the values of a_{YM}, induced by larger fluctuations due to the additional noise or to gossips.

4.2. *Biases with additional noise*

To check more deeply the validity of this analysis, we computed the biases of section 3.1 when adding an independent noise to a_{YM} when computing the bias on the opinion about others. The bias on a_{MM} remains the same as previously because the gossips do not change (directly) the dynamics on the self-opinion. This independent noise consists in adding randomly -0.04 or +0.04 to the changes of opinions. Figure 12 shows that the amplitude of the negative bias increases significantly for low fixed values of a_{MM}. This can be understood because the attractive effect to the anchor value of a_{YM} is small, which increases very significantly the fluctuations and thus the bias. As a result, the sum of the biases is always negative and more particularly

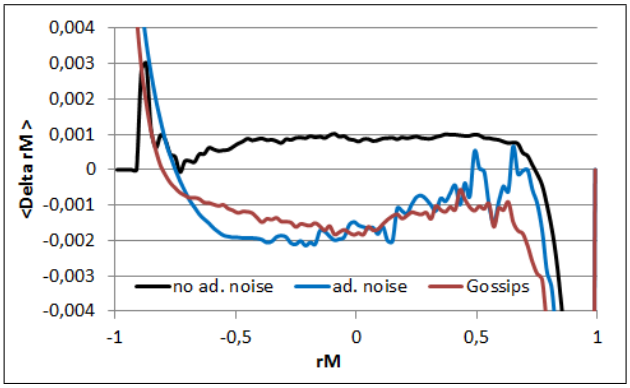


Fig. 11. Average variation of the time window averaged reputation $< \Delta \bar{r}_M >_i$ for 50 million iterations and 100 intervals. $a_{YY} \in \{-0.6, -0.4, -0.2, 0.0, 0.2, 0.4, 0.6, 0.8\}$, for the model with additional noise of ± 0.04, the model without this noise, and the model with gossips.

for negative values of a_{MM}. This should be compared with Figure 5, in which the positive bias is larger for negative values of a_{MM}.

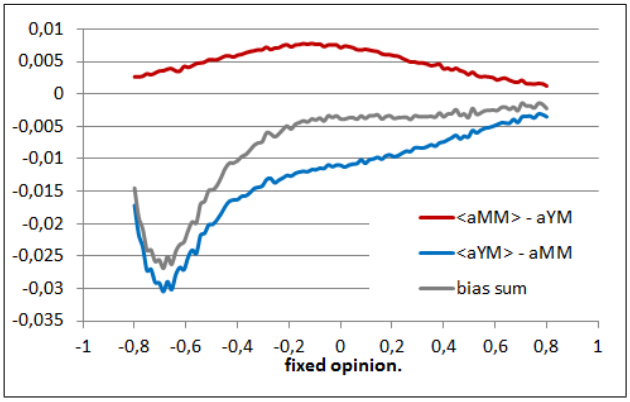


Fig. 12. Positive bias on self-opinion and negative bias on opinion about other when adding an independent noise of ± 0.04. The grey curve is the sum of both biases.

Figure 13 provides the results of running this model with an additional noise on the experiment of section 3.2, for two agents and fixing the values of $a_{YY} = a_{MY} = 0.0$. We observe that, with the additional noise, $< \Delta \bar{a}_{YM}(t) >_i$ is negative for $\bar{a}_{YM}$ higher than -0.5 whereas without the additional noise, $< \Delta \bar{a}_{YM}(t) >_i$ is positive for $\bar{a}_{YM}$ lower than 0.0. Therefore, with the additional noise, $\bar{a}_{YM}$ has a general downward tendency for a larger set of values than without this additional noise.

Finally, the analysis suggests that gossips increase the fluctuations on the opinions about others, which increase the negative bias about others, and particularly about agents with a low reputation (or self-opinion). This negative bias may become

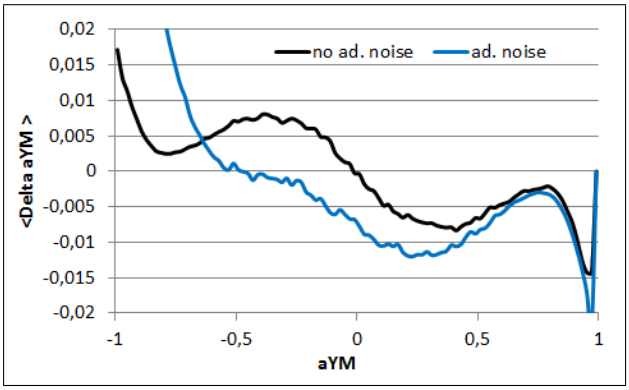


Fig. 13. Average variation of the time window averaged opinion $< \Delta \bar{a}_{YM}(t) >_i$ computed on 50 Million interactions between two agents M and Y, $a_{YY} = a_{MY} = 0.0$ fixed, for the model with additional noise of ± 0.04, and the model without this noise, for 100 intervals of $\bar{a}_{YM}(t)$.

stronger than the positive bias on the self-opinion and then it generates a downward tendency of the average opinion.

5. Discussion

Before considering possible interpretations of the model in social phenomena, it is important to discuss the robustness of our results with respect to our modelling choices.

- The attractive dynamics, with an intensity of attraction growing with the value given to the influencing agent, is well supported by many observations. However, some models of opinion dynamics make more refined assumptions (see [12] for a review) that could change our conclusions. For instance, the bounded confidence model [5, 18] which can be related to the confirmation bias, assumes that the opinion is not influenced by opinions which are too distant. Our analysis would not hold for interactions beyond the bounds, but it would for interactions within the bounds. Hence this change could modify partly our conclusions.
- Choosing a synchronous update leads to stronger positive and negative drifts than choosing an asynchronous update. This difference could be the subject of future research.
- The sigmoïd propagation function is probably a less common assumption. What is the impact of this choice on the model behaviour? Actually, the sufficient condition to produce the upward reputation variation (without gossips) is that the propagation function is growing with $(a_{MY} - a_{MM})$ and remains positive (see section 3.1). The effect should even be enhanced with a function that continues to grow more steadily than the sigmoïd function for positive values.
- The effect of gossips is only related to the importance of the sender, whereas it is independent from the value that the receiver gives to the target of the gossip.

Variants of the model taking the importance of the target of the gossip could be imagined, and their results might be different.

- In the model, positive and negative gossips are equally likely, which is not realistic. However, this hypothesis allows us to identify a statistical asymmetry and it would be straight-forward to investigate the effect of biased gossips in the model in the future.
- Starting with agents that are all identical is unrealistic, of course. However, as stressed in the introduction, this level of generality allows us to isolate the effect of the interactions alone, without any influence of intrinsic differences between the agents. Hence we assume that the properties of the interactions that are observed on the simplified dynamics are likely to also be present, though less easily identifiable, in more complicated and more realistic settings which could vary with specific properties of the agents. Of course, this assumption should be checked on more complicated models and, if possible, via human experiments.

Another perspective to discuss the relevance and the robustness of the model, is to notice that its conclusions are founded on two main mechanisms deriving from the modelling choices:

- When an agent increases her self-opinion, she tends to "go back" (decrease her self-opinion) less easily than before increasing it;
- When an agent decreases her evaluation of another agent, then this other agent has more difficulty to gain the esteem than before this decrease.

These mechanisms are indeed responsible for the effect of the fluctuations that are at the heart of the patterns that emerge in the model. Their validity can be discussed of course, but we argue that they seem to match standard common social experience. This work can be summarized as isolating these basic mechanisms and observing their statistical consequences on very large series of interactions from simulations.

Keeping in mind these strengths and weaknesses of the model, we turn now to the potential interest of its results in the study of real social dynamics. We propose three main subjects:

- The positive bias on self-opinions observed in the model could be the subject of an experimental research to check if it is observable in humans and if it can be related to the positivity bias identified by psychologists. Our approach could be considered as an alternative to the evolutionary explanation [23] of this human trait. Of course, this is the same for the negative bias about others.
- Our results suggest that a smooth leadership, namely feed-back opinions of the leaders about their subordinates which evolve slowly and cautiously, tends to help the subordinates to increase their self-opinion, for deep statistical reasons. Of course, in this case, other processes, for instance due to the public and official character of the hierarchy, take place and may overcome the processes that we

identified. Nevertheless, if the basic mechanisms of noisy interpretation and higher influence of higher valued are present, the identified biases should also be present.

- Gossips, though neutral on average by themselves in the model (because the noise is symmetric), appear to have an intrinsically negative effect on the opinions within the group for the same statistical reason. Therefore, leaders should be very cautious before being convinced, even by someone they trust, by a bad opinion about a member of their team. More generally, in the digital age which gives the possibility to spread gossips more easily than ever, it seems particularly important be aware of their potentially destructive effects and to develop protections from them. A better understanding of the mechanisms generating these effects could help in this endeavour.

6. Acknowledgement

This work has been partly supported by the Flagera project FuturICT2.0.

References

[1] Bagnoli, F., Carletti, T., Fanelli, D., Guarino, A. & Guazzini, A. 2007. Dynamical affinity in opinion dynamics modeling. *Phys. Rev. E* 76, 066105.

[2] Bonabeau, E., Théraulaz, G., Deneubourg, J.L. 1996. Mathematical model of self-organizing hierarchies in animal societies. *Bulletin of Mathematical biology*, 58(4), 661-717.

[3] Carletti, T., Fanelli, D. & Simone, R. 2011. Emerging structures in social networks guided by opinions' exchanges. *Advances in Complex Systems* 14, 28.

[4] Cox BA. 1970. What is Hopi gossip about? Information management and Hopi factions. *Man.* 5: 8898.

[5] Deffuant, G., D. Neau, F. Amblard, and G. Weisbuch. 2000. Mixing Beliefs among Interacting Agents. *Advances in Complex Systems* 3(01n04):8798

[6] Deffuant, G., Weisbuch, G., Amblard, F., Faure, T. 2003. Simple is beautiful... and necessary. *Journal of Artificial Societies and Social Simulation* 6 (1).

[7] Deffuant, G. 2006. Comparing Extremism Propagation Patterns in Continuous Opinion Models. *Journal of Artificial Societies and Social Simulation* 9(3):8.

[8] Deffuant, G., Moss, S., Jager, W. 2006. Dialogues concerning a (possibly) new science. *Journal of Artificial Societies and Social Simulation* 9 (1).

[9] Deffuant, G., Carletti, T., Huet, S. 2013. The Leviathan model: Absolute dominance, generalised distrust and other patterns emerging from combining vanity with opinion propagation. *Journal of Artificial Societies and Social Simulation* 16, 23

[10] Emler, N. 1990. A social psychology of reputation, *European Review of Social Psychology*, 1, 1, 171-193.

[11] Emler N. 2001. Gossiping. In: Giles H and Robinson WP (eds) *Handbook of language and social psychology.* Chichester: John Wiley & Sons, pp.317338.

[12] Flache, A., Msa, M., Feliciani, T., Chattoe-Brown, E., Deffuant, G., Huet, S., Lorenz, J. 2017. Models of Social Influence: Towards the Next Frontiers. *Journal of Artificial Societies and Social Simulation* 20 (4) 2.

[13] Flache, A. and M. W. Macy. 2011. Small Worlds and Cultural Polarization. *The Journal of Mathematical Sociology* 35(13):14676.

[14] French, J. R. 1956. A Formal Theory of Social Power. *Psychological Review* 63(3):18194.

[15] Foster, E. 2004.Research on gossip: Taxonomy, methods, and future directions.*Review of General Psychology*, 8,2, 78-99.
[16] Galam, S. 2002. Minority Opinion Spreading in Random Geometry. *The European Physical Journal B* 25(4):4036.
[17] Giardini, F., Conte, R. 2012. Gossip for social control in natural and artificial societies. *Simulation* 88(1):18-32.
[18] Hegselmann, R. and U. Krause. 2002.Opinion Dynamics and Bounded Confidence: Models, Analysis and Simulation. *Journal of Artificial Societies and Social Simulation* 5(3).
[19] Hoorens, V. 1993. Self-enhancement and superiority biases in social comparison. *European Review of Social Psychology* 4(1), 113-139.
[20] Hubbell, S.P. 2001. The Unified Neutral Theory of Biodiversity and Biogeography. *Princeton University Press.*
[21] Huet, S. and Deffuant, G. 2017. The Leviathan model without gossips and vanity: the richness of influence based on perceived hierarchy. *Advances in Social Simulation* Springer, Cham, 149-162.
[22] Huet, S., Edwards, M., Deffuant, G. 2017. *Taking into account the variations of social network in the mean-field approximation of the threshold behaviour diffusion model. J. Artif. Soc. Soc. Simulat.* 10.
[23] Johnson, D. and Fawler, J. 2011. *The evolution of overconfidence. Nature* 477:317-320.
[24] Huet, S. 2017. *The social functions of gossip and the Leviathan model.* in *Understanding Interactions in Complex Systems: Toward a Science of Interaction* 137.
[25] Lorenz, J. 2007. Continuous Opinion Dynamics under Bounded Confidence: A Survey. *International Journal of Modern Physics C* 18(12):181938.
[26] Panchanathan K and Boyd R. 2004. Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature.* 432: 499502.
[27] Piazza JR and Bering JM. 2008. Concerns about reputation via gossip promote generous allocations in an economic game. *Evol Hum Behav*, 6: 487501.
[28] Pornpitakpan, C.2004. The persuasiveness of source credibility: a critical review of five decades 321 evidence. *J. Appl. Soc. Psychol.* 34, 243281.
[29] Puge-Gonzalez, I., Hoscheid, A., Hemelrijk, C. 2015. Friendship, reciprocation, and interchange in an individual based model. *Behav. Ecol. Sociobiol.*. 69:383-394.
[30] Sabater J and Sierra C. 2005. Review on computational trust and reputation models. *Artif Intell Rev.* 24: 3360.
[31] Wert, S.R. and Salovey, P. 2004. A social comparison account of gossip, *Review of General Psychology*, 8, 2, 122-137.