THE JOURNAL OF MATHEMATICAL SOCIOLOGY
2022, VOL. 46, NO. 4, 315–341
https://doi.org/10.1080/0022250X.2021.1929968

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Hiding opinions by minimizing disclosed information: an obfuscation-based opinion dynamics model

Tanzhe Tang, Amineh Ghorbani, and Caspar G. Chorus

Department of Engineering Systems and Services, Faculty of Technology, Policy and Management, Delft University of Technology, Delft, The Netherlands

ABSTRACT
In the field of opinion dynamics, the hiding of opinions is routinely modeled as staying silent. However, staying silent is not always feasible. In situations where opinions are indirectly expressed by one's observable actions, people may however try to hide their opinions via a more complex and intelligent strategy called obfuscation, which minimizes the information disclosed to others. This study proposes a formal opinion dynamics model to study the hitherto unexplored effect of obfuscation on public opinion formation based on the recently developed Action-Opinion Inference Model. For illustration purposes, we use our model to simulate two cases with different levels of complexity, highlighting that the effect of obfuscation largely depends on the subtle relations between actions and opinions.

ARTICLE HISTORY
Received 28 September 2020
Revised 23 March 2021
Accepted 11 May 2021

KEYWORDS
Obfuscation; opinion-hiding; opinion dynamics; agent-based modeling; social simulation

1. Introduction

For diverse reasons, people may be unwilling to disclose their opinions to the public, especially when the topic is controversial. Instead, they may try to hide their opinions by adopting various strategies. As a prevalent behavior, hiding opinions has been studied in a number of opinion dynamics models. The majority of them assume that individuals hide their opinions by simply keeping silent (e.g., Gawronski et al., 2014; Ross et al., 2019; Sohn, 2019; Sohn & Geidner, 2016; Takeuchi et al., 2015). Although silence may help hide opinions from hearers of our words, it may not hide them from observers of our actions. That is, we usually learn others' opinions by inferring them from their actions based on some universal knowledge about how opinions and actions relate (Tang & Chorus, 2019). In other words, in daily life, actions are known to signal opinions, and as such, keeping completely silent is no longer feasible when observers learn opinions by observing actions.[1] For example, suppose a group of friends containing both vegetarians and

CONTACT Tanzhe Tang t.tang-1@tudelft.nl Department of Engineering Systems and Services, Faculty of Technology, Policy and Management, Delft University of Technology, Jaffalaan 5, 2628 BX Delft, The Netherlands

[1] This conclusion is still valid even if we generalize "keeping silent" to "doing nothing": although remaining quite in a debate is effortless, eating nothing in a dinner party seems less practical.

omnivores dine in a restaurant where only two dishes are available: beef steak and vegetable salad. If an omnivore wants to hide the opinion that it is OK to eat meat, choosing steak is certainly a bad idea, but keeping silent is also impractical. A better strategy is to choose salad, because both vegetarians and omnivores are more or less likely to eat salad, but only omnivores will order steak. Such a strategy, where actions are chosen that provide minimal information about underlying beliefs and preferences, is called obfuscation; it can be conceived as a manner to minimize the information disclosed to others by producing ambiguity and uncertainty (Chorus et al., 2020). Obfuscation has been a popular topic in software engineering (You & Yim, 2010) and more recently in privacy research (Brunton & Nissenbaum, 2015), but has not yet attracted attention in the community of opinion dynamics. When obfuscation behaviors are prevalent, a failure to capture them in models of opinion dynamics could lead to a biased understanding of how hiding opinions affects public opinion formation.

In this paper, we present an obfuscation-based opinion dynamics model to study the role of obfuscation in public opinion formation by embedding the obfuscation mechanism (Chorus et al., 2020) into the Action-Opinion Inference (AOI) modeling framework (Tang & Chorus, 2019), where people choose actions according to their opinions and learn others' opinions by interpreting their actions. Within this AOI framework, an obfuscating individual would hide her opinion by choosing the action that (i) is permitted by or in line with her opinion, yet (ii) releases the least amount of information about the opinion to others. This model fills the gap between existing models where hiding opinions equates to keeping silent and the reality that the mechanism of hiding opinions can be more subtle and complex than simply staying silent. As a result, our model can offer a more realistic and reasonable explanation of various social phenomena related to public opinion formation, particularly on (morally) sensitive topics. For example, incorporating obfuscation in the model can lead people to overestimate the popularity of the opinion that obliges the observed action. In a relatively simple setting, this may result in a significantly larger population believing in this opinion, which would have otherwise been different if only "silent-keeping" was considered.

The remainder of the paper is organized as follows: In Section 2, we review existing opinion dynamics models of hiding opinions and explain the foundation of our model. Section 3 describes the model in detail. In Section 4, we provide two illustrative examples abstracted from daily life and tales, in order to illustrate how this model works. Section 5 provides a brief summary and outlooks for further research.

2. Theoretical background

2.1. Hiding opinions in opinion dynamics

Opinion dynamics is one of the most popular and well-established fields in sociophysics. By modeling how opinions spread between individuals at a micro level, opinion dynamics models aim to explain macro-level phenomena such as polarization and consensus in a group of interacting individuals. Most opinion dynamics models pay little attention to the notion that people might want to hide their opinions and routinely assume that opinions can be directly observed, and that individuals always express opinions honestly (Mitsutsuji & Yamakage, 2020; Tang & Chorus, 2019). This assumption is likely to be unrealistic in circumstances where opinions are not completely visible, or individuals want to hide their opinions to avoid shame or to protect their privacy more generally.

Recently, however, a number of models[2] involving opinion-hiding have been proposed. The majority of them are based on the so-called spiral of silence theory, postulating that due to the fear of social isolation, people are more likely to keep silent if they think they are in the minority (Noelle-Neumann, 1974). In spiral of silence models, the choice between keeping silent and expressing one's opinion is determined by individual's perception of the opinions of others. For example, Gawronski et al. (2014) assume that the probability of expressing opinions is a negative function of the absolute difference between individual's own opinion and her perceived public opinion. Others prefer a threshold rule: in Sohn and Geidner's model (Sohn & Geidner, 2016), as well as a more recent one (Sohn, 2019), an individual speaks out if the intensity of her opinion is larger than the expression threshold, which is a personal and constant attribute. Following this tradition, Ross et al. (2019) introduce a similar attribute called willingness to self-censor. The condition of speaking out is that an individual's confidence in her opinion is larger than her willingness to self-censor, and the level of confidence is positively related to the proportional difference between the number of neighbors who agree and disagree with the individual.

Other models of hiding opinions follow different theories. For example, Grandi et al. (2017) consider hiding or disclosing opinions as a strategy to achieve a certain goal by influencing others' opinions. Fan and Pedrycz (2015, 2016) adopt the social judgment theory and postulate that people remain silent if the intensity of their preference for one of two alternatives is not strong enough. As a conclusion, most models involving the behavior of hiding opinions, regardless of their theoretical basis, take it for granted that hiding opinions equates to keeping silent.

[2] In some of these models (e.g., Gawronski et al., 2014 & Ross et al., 2019), opinions are fixed, and agents update their choices between expressing opinions and keeping silent. We regard them as an extended class of opinion dynamics models.

2.2. Obfuscation and action-opinion inference

Opinions are not always expressed by words but can also be revealed by actions. As argued in Section 1, when observers learn someone's opinion by observing her actions, keeping silent is not (always) possible; we claim that in such a case, obfuscation becomes the best strategy.

In the past few decades, most obfuscation studies were conducted in the computer science domain, especially software engineering, where code obfuscation is a very popular topic (You & Yim, 2010). More recently, philosophers and social scientists started to pay attention to obfuscation with a special interest in how obfuscation can be used to defend one's privacy on the Internet (Brunton & Nissenbaum, 2015; Davis, 2019; Doyle, 2018), and how obfuscation mitigates unfavorable moral reactions to morally disreputable economic exchanges (Rossman, 2014; Schilke & Rossman, 2018; Wherry et al., 2019).

In the context of a coordination problem, Dewan and Myatt (2008) consider obfuscation as a technique of a leader to compete for audience by deliberately reducing the clarity of her message. Technically, obfuscation is modeled by manipulating "the variance of the noise in her speech" (Dewan & Myatt, 2008). In game theory, obfuscation is most closely related to intentional vagueness, i.e., deliberately choosing vague messages even if more precise alternatives are available (Blume & Board, 2014). A number of studies present the "game-theoretic rationale for vagueness" by showing that vagueness can "mitigate conflict" and "enhance efficiency" in a sender-receiver game (Blume & Board, 2014; De Jaegher, 2003; Serra-Garcia et al., 2011). Both Dewan-Myatt's obfuscation and intentional vagueness mainly deal with verbal communications, and their analyses are often based on calculations of utilities. In this paper, we embed obfuscation in non-verbal communications where opinions are signaled by actions, and we are more interested in the effect of obfuscation on opinion dynamics rather than people's utility or the equilibrium of a particular game.

In recent years, the concept of obfuscation has been introduced and formalized as a communication strategy in choice modeling. Chorus et al. (2020) combine the notions of Bayesian inference and Shannon entropy, integrating them into a formal model of obfuscation-based decision-making. The idea is that a subject knows that her actions signal her underlying preferences (opinions) and selects the action that is in line with her preferences while providing as little as possible information to observers. The model is designed to describe the behaviors of humans whose actions are observed by others as well as the behaviors of autonomous agents under the surveillance of a human supervisor. In the model, agents choose actions based on a particular rule (here: opinion) that is unknown to the supervisor. Based on the observation of the agent's action, the supervisor infers the opinion that motivates the action according to the Bayes' Theorem. An obfuscating human or autonomous agent, being aware that it might be "punished" if the observer or supervisor learns that it has an "unwanted" opinion, will choose actions by maximizing the Shannon entropy generated by its choice while staying as close as possible to its opinion.

To utilize this mechanism in the context of opinion dynamics, we first need to formalize how opinions are learned by observing actions. In fact, such a formalization exists in the form of a so-called Action-Opinion Inference (AOI) model (Tang & Chorus, 2019). In the AOI model, the relation between opinions and actions is described by deontic logic: an opinion can oblige, permit, or prohibit an action. Equipped with the action-opinion relation, individuals "infer the opinions of others by observing and interpreting their actions" (Tang & Chorus, 2019). Based on the inference, individuals update their own opinions "according to the relative probability of each opinion in the neighborhood, calculated from the inferences of different opinions" (Tang & Chorus, 2019). As the final step, individuals choose new actions according to the newly updated opinions. The AOI model is compatible with Chorus et al.'s obfuscation mechanism not only because it formalizes the notion of "learning opinions by observing actions", but also because of the deontic logic underlying the action-opinion relation, where an action may be driven by different opinions, and an opinion may permit different actions, allowing agents to obfuscate by choosing certain actions. If each action is driven by only one opinion, observers can then directly and correctly read opinions from actions, and there will be no room for obfuscation.

At the end of this section, we would like to point out the connection between the AOI model and social learning models in economics. Both types of models study how people infer (learn) and aggregate opinions (information) from their social environment (Golub & Sadler, 2016). In the AOI model, agents update their opinions in a Bayesian manner, which is a common setting in social learning (the so-called "Bayesian social learning", e.g., Acemoglu et al., 2011; Gale & Kariv, 2003). The AOI model is also closely related to "observational social learning" in which agents observe choices made by their predecessors (Çelen & Kariv, 2004). Despite these similarities, the AOI model highlights the multiplicity of action-opinion relations, while social learning models may pay more attention to convergence and efficiency (Golub & Jackson, 2010; Lobel et al., 2009; Mossel et al., 2016). In particular, social learning models have a constant interest in convergence to the true/accurate opinion (Golub & Jackson, 2010; Jadbabaie et al., 2012) or the right/best action (Acemoglu et al., 2011) via learning, but the AOI model (or opinion dynamics models in general) does not involve any judgment or evaluation. We therefore conclude that the AOI model is located at the boundary (which itself is blurred) between opinion dynamics and social learning, and hence our obfuscation model – whose basis is the AOI model – relates to both disciplines other than opinion dynamics alone.

3. The model

In this section, we develop an opinion dynamics model of obfuscation by embedding the obfuscation mechanism (Chorus et al., 2020) in the framework of the Action-Opinion Inference (AOI) model (Tang & Chorus, 2019).

The basic model setup resembles the AOI model. We consider a population of N agents located on an undirected network G that describes how agents are connected. Agents are neighbors if they are directly connected in the network. Each agent i ($i = 1, 2, \ldots, N$) holds an invisible opinion $o^{(i)}$ from the opinion set $O = \{o_1, \ldots, o_k, \ldots, o_K\}$, based on which she chooses a visible action $a^{(i)}$ from the action set $A = \{a_1, \ldots, a_g, \ldots, a_G\}$. The relation between o_k and a_g is denoted by $s_{kg} \in \{\pm 1, 0\}$, where $s_{kg} = 1$ implies a_g is obliged by o_k, $s_{kg} = 0$ implies a_g is permitted by o_k, and $s_{kg} = -1$ implies a_g is forbidden by o_k. All s_{kg} ($k = 1, \ldots, K; g = 1, \ldots, G$) compose the so-called action-opinion matrix $S = \{s_{kg}\}$. Agents are assumed to have the same action set, opinion set, and action-opinion matrix. This assumption will be relaxed in Section 4.2, where people may have difference perceptions of the relation between actions and opinions.

Assume that there is a fixed number of N_o obfuscators in the population who want to hide their opinions, and $N - N_o$ non-obfuscators who do not care if their opinions are disclosed or not. Initially (i.e., *stage 0*), each agent (both obfuscators and non-obfuscators) is randomly assigned an opinion from the opinion set O, based on which she chooses an action from the action set A according to the rule of updating actions (the rule will be given in Section 3.1 and 3.2).

In each time step, an agent, whether an obfuscator or not, is randomly chosen to go through the following successive stages: (1) *observing actions and inferring opinions*, (2) *updating opinions*, and (3) *updating actions*. For the sake of clarity, we will demonstrate the behaviors of obfuscators and non-obfuscators separately.

3.1. Behavior of non-obfuscators

(0) choosing actions based on the initial opinions

Before any agent is chosen to go through the three main stages, each agent needs to choose an action based on her initial opinion. The rule of choosing actions of a non-obfuscator is as follows: if the opinion of a non-obfuscator i, $o^{(i)} = o_k$, obliges an action a_g, she will certainly choose this action because it is the only option. Formally, the probability of choosing a_g when holding o_k, $P(a_g|o_k)$, equals 1 if $s_{kg} = 1$. If $o^{(i)} = o_k$ forbids a_g, agent i will not choose a_g. That is, $P(a_g|o_k) = 0$ if $s_{kg} = -1$. If $o^{(i)} = o_k$ permits more than one action, agent i will choose one of these permitted actions with equal probability.

Formally, $P(a_g|o_k) = \frac{1}{W}$ if $s_{kg} = 0$, and W is the number of actions permitted by o_k. To summarize:

$$P(a_g|o_k) = \begin{cases} 1 & if\ s_{kg} = 1 \\ 0 & if\ s_{kg} = -1 \\ \frac{1}{W} & if\ s_{kg} = 0 \end{cases} \tag{1}$$

(1) observing actions and inferring opinions

Once an agent is chosen, she first observes the actions chosen by her neighbors, based on which she infers neighbors' opinions behind these actions. After observing neighbor j choosing action $a^{(j)}$, agent i believes that the opinion of j is o_k with probability $P^{(i)}(o^{(j)} = o_k|a^{(j)})$, which takes the following form:

$$P^{(i)}\left(o^{(j)} = o_k|a^{(j)}\right) = \frac{P(a^{(j)}|o_k)}{\sum_{z=1}^{K} P(a^{(j)}|o_z)} \tag{2}$$

where $P(a^{(j)}|o_z)$ is the probability of choosing $a^{(j)}$ when holding opinion o_z, and can be calculated by Eq. (1). We can derive Eq. (2) from the Bayes' rule by assuming the prior probability $P(o_z) = \frac{1}{K}$ for all $z = 1, 2, \ldots, K$. The rationale behind this assumption is that agents have no prior knowledge about which opinion is more likely to be adopted by their neighbors before observing their actions.

(2) updating opinions

After inferring the opinions of all neighbors, agent i evaluates the relative probability of each opinion in the neighborhood:

$$\hat{P}^{(i)}(o_k) = \frac{\sum_{j\in M_i} P^{(i)}(o^{(j)} = o_k|a^{(j)})}{\sum_{z=1}^{K}\sum_{j\in M_i} P^{(i)}(o^{(j)} = o_z|a^{(j)})}, k = 1, 2, \ldots, K \tag{3}$$

where M_i is the set of all agent i's neighbors. As a result of positive social influence (Flache et al., 2017), agent i will update her opinion to o_k with probability $\hat{P}^{(i)}(o_k)$. In case that other forms of social influence or mechanism are preferred, modelers can easily modify Eq. (3) accordingly.

(3) updating actions

In the last stage, the chosen agent updates her action based on her opinion that has just been updated in the previous stage. This stage follows the same rule as in stage 0 where non-obfuscators choose their actions based on their initial opinions. Then, the world goes to the next time unit.

To summarize, for a chosen agent, one time unit includes all the three stages: *observing actions and inferring opinions*, *updating opinions*, and *updating* *actions*. We define a time step as N successive time units, therefore on average in a time step everyone has one chance to update her opinion and action.

3.2. Behavior of obfuscators

The behavior of an obfuscator is the same as a non-obfuscator in stage 1 and 2. The only difference lies in the rule of choosing actions, which applies to both stage 0 and 3. First of all, an obfuscator is still governed by the action-opinion relation: she must choose the obliged action and cannot choose the forbidden action. As a result, an obfuscator can only play obfuscation when her opinion permits more than one action. Among all the actions that are permitted by her opinion, according to Chorus et al. (2020), an obfuscator chooses the (permitted) action that reveals as little information as possible about the opinion by maximizing the uncertainty of her decision, measured by the Shannon entropy. For each action a_g, the Shannon entropy is calculated by:

$$H\left(a_g\right) = -\sum_{k=1}^{K} P\left(o_k|a_g\right)\log\left(P\left(o_k|a_g\right)\right) \tag{4}$$

where $P\left(o_k|a_g\right)$ is short for $P^{(i)}\left(o^{(j)} = o_k|a^{(j)} = a_g\right)$, thus it can be calculated by Eq. (2). Larger entropy implies more uncertainty. If $H\left(a_g\right) = 0$ (i.e., the entropy is minimized), choosing a_g reveals the full amount of information regarding the invisible opinion. To support this claim, we must show that $H\left(a_g\right) = 0$ only if there exists a $k = k^*$ such that $P\left(o_{k^*}|a_g\right) = 1$, and $P\left(o_k|a_g\right) = 0$ for all $k \neq k^*$. Fortunately, this has been proven by Shannon (1948) as a basic property of the Shannon entropy. Meanwhile, the entropy is maximized, according to Eq. (4), when $P\left(o_m|a_g\right) = P\left(o_n|a_g\right)$ for all $m, n = 1, \ldots, K$, that is, the observer has no knowledge about which opinion is more likely to be the opinion of an agent choosing a_g. In practice, this perfect maximization is not always achievable due to the restriction of the action-opinion matrix.

Formally, an obfuscator i will choose $a^{(i)}$ according to:

$$a^{(i)} = \arg max_{a_g \in A_i} H\left(a_g\right) \tag{5}$$

where A_i is the set of actions available to i. In other words, A_i contains all the actions permitted or obliged by $o^{(i)}$.

It is worth noting that both obfuscators and non-obfuscators know nothing about the identities (i.e., obfuscator or non-obfuscator) of their neighbors, nor do they know the number of obfuscators in the population. The assumption can be relaxed if modelers want to study more intelligent agents who are able to learn the identities of others.

Figure 1 gives a brief summary of the model. Firstly, each agent is randomly assigned an opinion. Then obfuscators and non-obfuscators choose actions

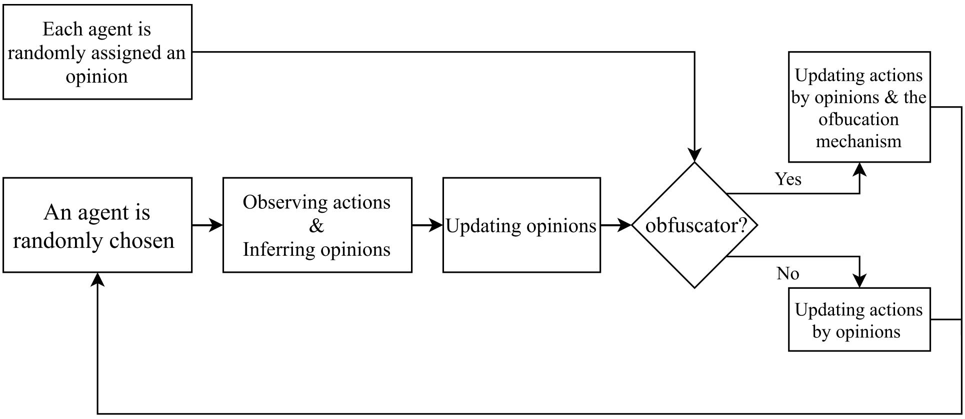

Each agent is randomly assigned an opinion
An agent is randomly chosen
Observing actions & Inferring opinions
Updating opinions
obfuscator?
Yes
No
Updating actions by opinions & the ofbucation mechanism
Updating actions by opinions

Figure 1. Illustration of the model.

based on different rules (stage 0). Afterward, a random agent is selected to update her opinion and action through a three-stage process: inferring opinions of others (stage 1), updating opinion based on the inference (stage 2), and updating action based on the updated opinion (stage 3).

4. Illustrative examples

As we will soon witness in this section, the effect of obfuscation on public opinion largely depends on the relations between actions and opinions. Understanding the effect of obfuscation in a particular case requires running simulations of the model under particular conditions. To illustrate how this works, we provide two examples. The first example that describes the dynamics of vegetarians and omnivores is extremely simple, aiming to provide a step-by-step demonstration. The second example, trying to explain the ironic situation in *The Emperor's New Clothes*, is more subtle and complex, as people with different opinions have different perceptions of the relation between actions and opinions. It is important to note here, that the sole aim of these examples is to illustrate the workings of the obfuscation-based opinion dynamics model – as such we refrain from drawing any generic (i.e., not specific to the example) conclusions about the potential effect of obfuscation on opinion dynamics. For that, a larger number of more elaborate case studies are needed which are preferably grounded in real life opinion formation situations.

4.1. The battle between vegetarians and omnivores

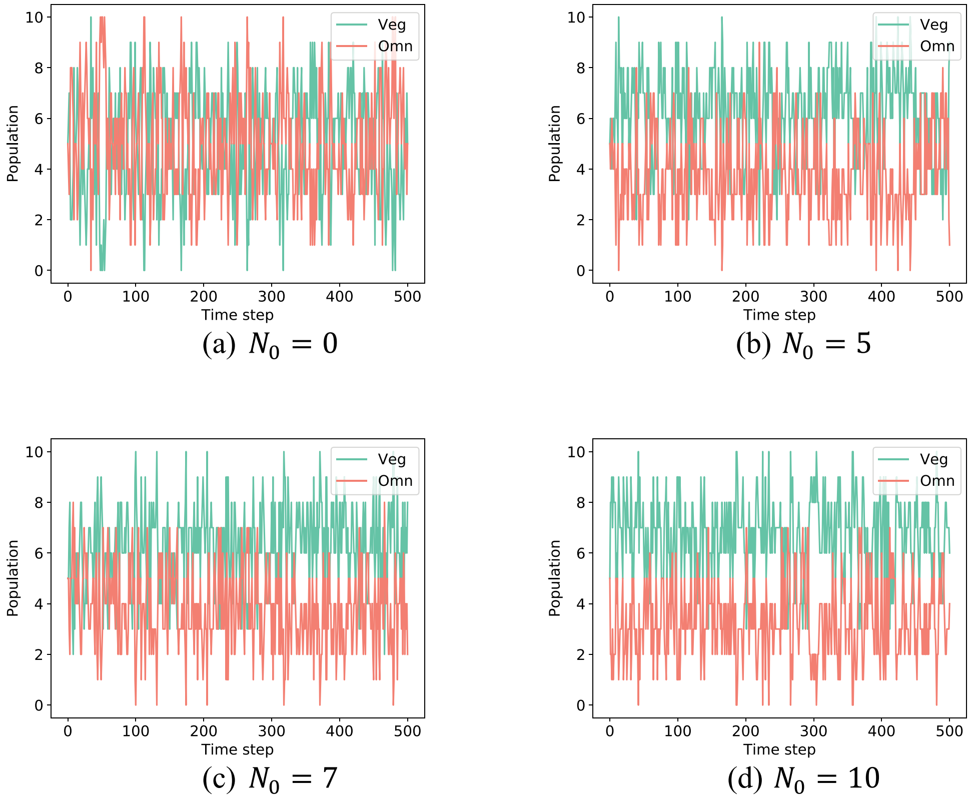
We first look into a very simple case, the vegetarian-omnivore example mentioned in Section 1. Here we assume that there are $N = 10$ friends going to the restaurant. Initially, $N_{Veg} = 5$ of them are vegetarians and $N_{omn} = 5$ of them are omnivores. Given the relatively small population, it is reasonable to assume that everyone can observe the action of everyone else. We summarize this case by the following parameters and conditions: $N = 10$, G is a complete graph (i.e., everyone is a neighbor of everyone else), $A = \{a_1 = (Choose)Steak, a_2 = (Choose)Salad\}$, $O = \{o_1 = Veg, o_2 = Omn\}$ ("*Veg*" is short for "Vegetarian", and "*Omn*" is short for "Omnivore"), and the action-opinion matrix S^{VO} ("*VO*" stands for "the battle between **V**egetarians and **O**mnivores"):

$$S^{VO} = \begin{array}{ccc} & Steak & Salad \\ Veg & -1 & +1 \\ Omn & 0 & 0 \end{array}$$

which means a vegetarian is prohibited from choosing steak and can only choose salad, while an omnivore can choose between steak and salad. Without any calculation, we can already see that an obfuscating omnivore will choose salad, and a non-obfuscating omnivore will choose randomly (i.e., flip a coin) between steak and salad. It is also worth noting that whether a person obfuscates does not depend on the chosen action. For example, an obfuscating vegetarian should make the same choice (i.e., salad) as a non-obfuscating vegetarian. However, the obfuscating vegetarian chooses salad because it gives the minimum information, while the non-obfuscating vegetarian makes the same choice because it is the only permitted option, regardless of how much information it releases. In practice they choose the same action, but their motivations for doing so are different.

The running time is set to be 500, which is sufficiently long to reach a stable outcome. Figure 2 shows how the number of obfuscators in the population affects public opinion, based on which we can conclude that obfuscation, in this particular case, suppresses the spread of omnivorism and promotes the popularity of vegetarianism. However, the effect is bounded: in Figure 2(d), even if everyone is an obfuscator, in equilibrium, there still exist a few omnivores (around 1 to 2), implying that obfuscation cannot completely eliminate the existence of omnivorism.

To further explore the relation between obfuscation and public opinion, we run the simulation 100 times for each N_o. In Figure 3, $\bar{f}_{Veg}$ (i.e., the y axis) is the fraction of vegetarians in the population averaged over the last 50 time steps of each simulation realization. Compared to Figure 2, Figure 3 provides more details. We can see that although obfuscation (represented by N_o) has a significant impact on public opinion (represented by $\bar{f}_{Veg}$), it is not a fully determining factor, as there remains a remarkable degree of variation across realizations regardless of N_o. This statement comes from the observation that even if all conditions and parameters (including N_o) are the same, the public opinion in each realization can be very different. For example, when $N_o = 0$, the lowest $\bar{f}_{Veg}$ is close to 0.35, and the highest is about 0.6. However, there is a trend that this variation decreases as the



(a) $N_0 = 0$

(b) $N_0 = 5$

(c) $N_0 = 7$

(d) $N_0 = 10$

Figure 2. The battle between vegetarians and omnivores: population of believers in each opinion versus time step.

number of obfuscators increases. As another interesting finding, we can conclude that obfuscation is likely to reduce the variation in public opinion for this particular case.

The rationale behind the simulation result lies in the discrepancy in observer's inference and the reality: from Eq. (2) and S^{VO}, we know that observers believe a salad-eating agent is an omnivore with a probability of $1/3$. However, because an obfuscating omnivore always choose salads, this probability is in fact larger than $1/3$ in a population containing obfuscators. In the extreme case where everyone is an obfuscator, the probability increases to $1/2$, the same as the probability that a salad-eating agent is a vegetarian. Such a discrepancy leads to an underestimation of the population of omnivores (or, equivalently, overestimation of the population of vegetarians). Consequently, vegetarianism becomes more popular than omnivorism because of positive social influence. At the same time, omnivorism will not go extinct because observers believe that omnivores are always likely to exist (with a relatively small probability) even if everyone chooses salads.

The rationale described above is formally expressed in the Appendix, where analytical results of this example are derived. According to the derivation, the

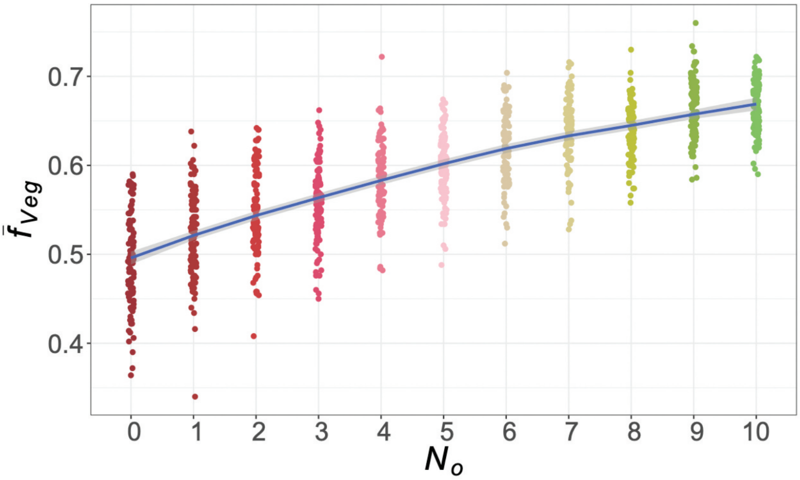


Figure 3. The battle between vegetarians and omnivores: fraction of vegetarians ($\bar{f}_{Veg}$) versus the number of obfuscators (N_o). For each N_o, we run 100 realizations of the simulation. $\bar{f}_{Veg}$ is obtained by averaging the fraction of vegetarians in the last 50 time steps of a realization. Each data point represents one realization. The horizontal position of each data point is slightly adjusted in order to reduce overlap. The line across the figure is the smoothed conditional mean, and the shaded area indicates the 95% confidence interval.

fraction of vegetarians, averaged over all realizations (trajectories) of the dynamics, should converge to $\frac{1+\theta}{2+\theta}$ over time, where $\theta = \frac{N_o}{N}$ is the fraction of obfuscators in the population. This conclusion is validated by the simulation result in Figure 3, where the average $\bar{f}_{Veg}$ (averaged over 100 independent realizations) is well approximated by $\frac{1+\theta}{2+\theta}$. Furthermore, the derivation shows that the average $\bar{f}_{Veg}$ only depends on the fraction of obfuscators (θ) and the action-opinion matrix (S^{VO}), while other conditions such as the size of the population (N) and the initial distribution of each opinion (N_{Veg} and N_{Omn}) are irrelevant. Unexpectedly, the number of neighbors of each agent is also irrelevant, as long as everyone has the same number of neighbors.[3]

Finally, we show that the effect of obfuscation on public opinion largely depends on the relations between actions and opinions. If we replace the omnivores here with carnivores (abbreviated to "*Car*") who only consume meats, the matrix is now S^{VC}("*VC*"stands for "The battle between **V**egetarians and **C**arnivores"):

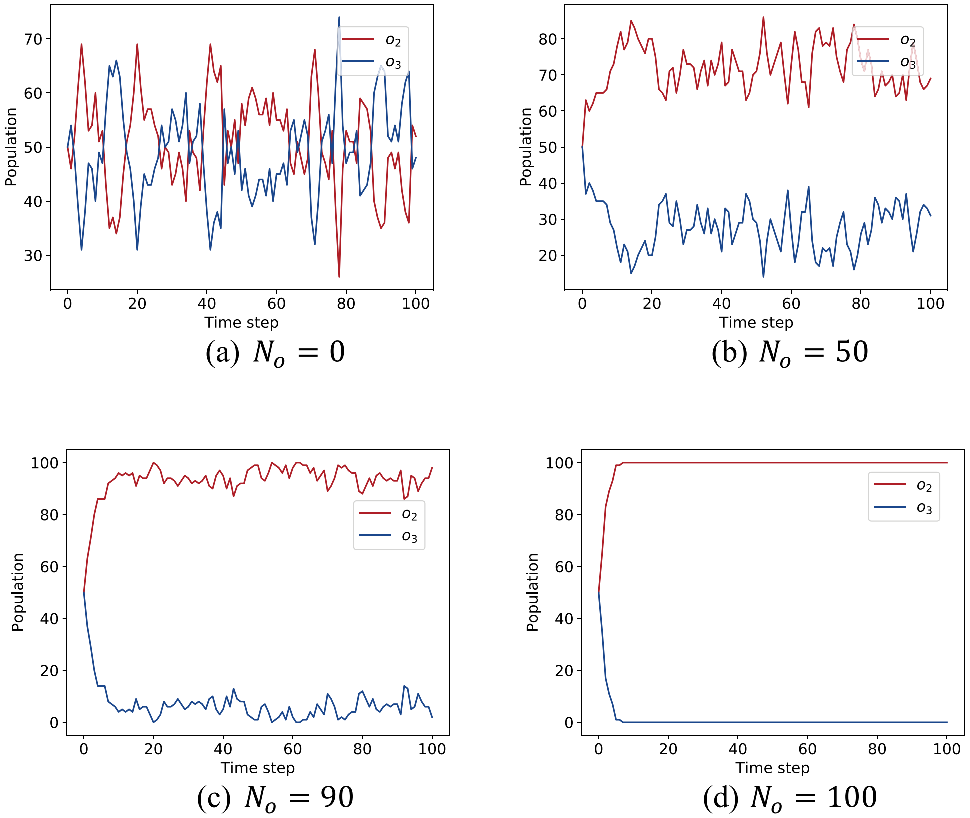
$$S^{VC} = \begin{array}{ccc} & Steak & Salad \\ Veg & -1 & +1 \\ Car & +1 & -1 \end{array}$$

[3] Readers should be aware that (1) the derivation benefits from the simplicity of S^{VO} and the assumption that everyone has the same number of neighbors; and (2) the conclusions made here are completely based on the derivation. It is unclear if they are valid in other cases.

It is obvious that obfuscation plays no role in this new case as vegetarians can only choose salad and carnivores can only choose steak. In other words, an obfuscator behaves the same as a non-obfuscator. The opinion dynamics described by S^{VC} has been solved analytically in the studies of the voter model (Krapivsky et al., 2010), from which we learn that the population would eventually reach a consensus of either vegetarianism or carnivorism (Tang & Chorus, 2019). As a result, the conclusion drawn from S^{VO} is invalid for the situation described by S^{VC}.

4.2. The emperor's new clothes

The Emperor's New Clothes is a famous tale written by Hans Christian Andersen in 1837. The general plot is about how two swindlers pretending to be weavers, convince the Emperor that the suit of clothes they made is invisible to stupid people. Everyone in the country, after observing the naked Emperor, out of fear of being considered stupid, pretends that they could see the clothes, until a child spoke out the truth.



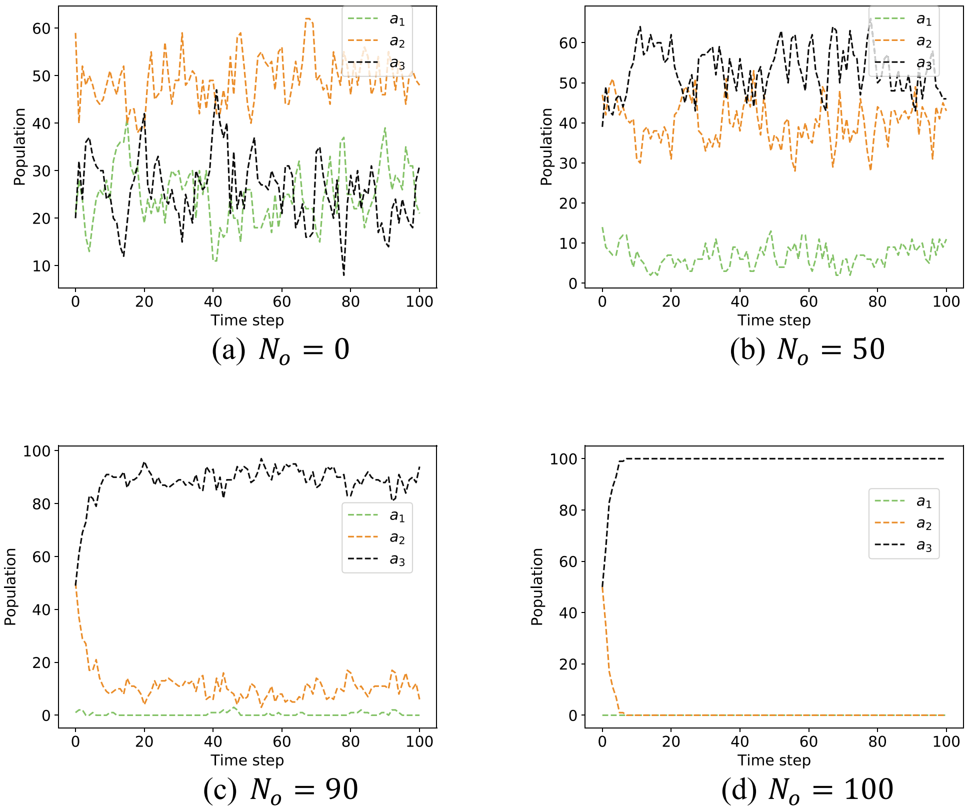
(a) $N_o = 0$

(b) $N_o = 50$

(c) $N_o = 90$

(d) $N_o = 100$

Figure 4. The emperor's new clothes: opinion dynamics of the citizens. Population of believers in o_1 is always zero and thus is not plotted.



(a) $N_o = 0$

(b) $N_o = 50$

(c) $N_o = 90$

(d) $N_o = 100$

Figure 5. The emperor's new clothes: population dynamics of actions chosen by the citizens versus time step.

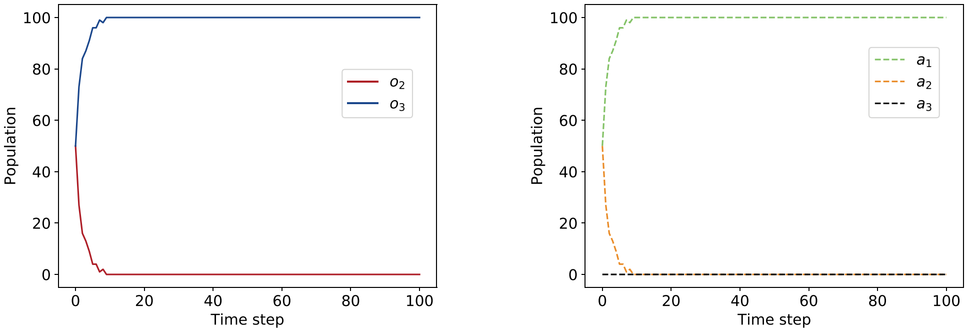
For sociologists, the tale, as a symbolic example of "support for a public lie" (Centola et al., 2005), is of particular interest because the ironic phenomenon that everyone pretends that they can see the clothes needs further explanation: besides the fear of being labeled stupid, is there any other mechanism underlying the phenomenon? One of the most popular explanations uses the concept of pluralistic ignorance (e.g., Bjerring et al., 2014; Centola et al., 2005; Hansen, 2012). Pluralistic ignorance describes a situation where most people privately reject or disapprove an opinion, but (incorrectly) believe that the opinion has been widely accepted by others (Miller & McFarland, 1987). To explain the tale by pluralistic ignorance, citizens in the tale are assumed to be "disbelievers" as they in fact think the Emperor is naked. Then the phenomenon is achieved when all disbelievers publicly praise the invisible suit based on the false belief that everyone else thinks the Emperor is not naked (Bjerring et al., 2014).

"Naked emperors are easy to find but hard to explain." (Centola et al., 2005). Despite being a popular practice to explain the tale, pluralistic ignorance overlooks the dynamics of opinions in the population. In fact, keeping one's opinion unchanged is a basic condition of pluralistic ignorance. For example, in Centola's model (Centola et al., 2005), the population is divided into "true believers" who always admire the Emperor, and "disbelievers" who (privately) think the Emperor is naked. Both true believers and disbelievers are not allowed to change their private opinions, regardless of their compliance decision.[4] This naturally raises the following questions: if individuals are allowed to change opinions, can this "public lie" become a "(false) public opinion" where everyone believes that the Emperor is dressed? In the other extreme, can this "public lie" become a "public truth" where everyone not only privately believes but also publicly claims that the Emperor is naked?

To answer these questions, we need to take an alternative approach. In the rest of the section, we will explain the tale from the perspective of opinion dynamics and obfuscation by investigating the role of obfuscation in the dynamics of opinions among citizens, including both "true believers" and "disbelievers".

While some citizens believe that the Emperor is naked, others may believe that the Emperor is dressed, and that they cannot see the clothes because they are stupid. Naturally, the latter will have the false imagination that some other citizens can see the clothes. To summarize, there will be three opinions involved in the story:

- o_1: I can see the clothes because I am not stupid.
- o_2: I cannot see the clothes because I am stupid.
- o_3: I cannot see the clothes because the Emperor is naked.



(a) Dynamics of opinions (b) Dynamics of actions

Figure 6. The emperor's new clothes: population dynamics of opinions and actions chosen by the citizens. every citizen is transparent.

[4] In the extension of Centola's model within the same paper, disbelievers with false enforcement are allowed to convert to true believers, but true believers cannot convert to disbelievers by default.

It should be noted that o_1 is imaginary, as in fact no one would hold this opinion.

Relevant actions, as one can imagine, include:

- a_1: publicly mock the Emperor/point out that the Emperor is naked.
- a_2: keep silent.
- a_3: publicly admire the Emperor's clothes.

Citizens who believe in o_2 perceive the following action-opinion relation:

$$S^{EC}(o_2) = \begin{array}{cccc} & a_1 & a_2 & a_3 \\ o_1 & -1 & -1 & +1 \\ o_2 & -1 & 0 & 0 \\ o_3 & 0 & 0 & -1 \end{array}$$

where "*EC*" stands for "**E**mperor's New **C**loth". Because citizens with $o_1$only exist in the imagination of citizens with o_2, the row that describes o_1 is completely determined by how citizens with o_2 think: without any social pressure, citizens with o_1 are expected (by citizens with o_2) to have no motivation to *mock the Emperor* (a_1) or *keep silent* (a_2). Meanwhile, citizens with o_2, although they cannot see the clothes, will never *mock the Emperor* (a_1) because they believe the clothes do exist. For citizens with o_3 who disbelieve the lie, citizens with o_2 assume that they would never *admire the Emperor* (a_3).

Citizens with o_2 are facing the (maybe imaginary) social pressure of being labeled as stupid people, and therefore have the incentive to hide their opinion by obfuscation. It seems that an obfuscator who believes in o_2 should choose a_2 over a_3 due to the fact that the entropy of a_2 is larger than that of a_3 according to $S^{EC}(o_2)$. However, in this particular case, citizens with o_2 believe that the pressure only comes from those who believe in o_1, because citizens with o_3, by definition, do not accept the swindlers' lie, hence they would not consider citizens with o_2 to be stupid. As a result, citizens with o_2 only care about the judgment from citizens with o_1, and will choose actions based on the action-opinion relation perceived by citizens with o_1 instead of their own perception $S^{EC}(o_2)$. Because citizens with o_1 only exist in the imagination of citizens with o_2, the perception of the action-opinion relation by citizens with o_1 is determined by citizens with o_2, and is therefore denoted by $S^{EC}(o_1|o_2)$:

$$S^{EC}(o_1|o_2) = \begin{array}{cccc} & a_1 & a_2 & a_3 \\ o_1 & -1 & -1 & +1 \\ o_2 & -1 & 0 & 0 \end{array}$$

The absence of o_3 is because we assume citizens with o_2 believe that citizens with o_1 would ignore the existence of o_3. The rationale behind this assumption is that citizens with o_1 might be so confident in their opinion that they do not expect others would think the Emperor is naked.[5] It is clear from $S^{EC}(o_1|o_2)$ that an obfuscator with o_2 would choose a_3, because choosing a_2 is a signal of being stupid in the eyes of citizens with o_1, as citizens with o_2 believe.

Observing more people choosing a_3 makes citizens with o_2 believe that there are more citizens with o_1 (i.e., $\hat{P}^{(i)}(o_1)$ increases, where i denotes citizens with o_2). However, they cannot change their opinion from o_2 to o_1, therefore observing a_3 only makes them more confident in their current opinion o_2.

Now, we consider citizens with o_3, the disbelievers. As they believe the Emperor is naked, to them, o_1 does not exist. Therefore, their perception of the action-opinion relation is[6]

$$S^{EC}(o_3) = \begin{array}{cccc} & a_1 & a_2 & a_3 \\ o_2 & -1 & 0 & 0 \\ o_3 & 0 & 0 & -1 \end{array}$$

Citizens with o_3 also have incentives to play obfuscation as they may not want to be considered stupid by citizens with o_2. To hide their opinion, instead of referring to their own perception $S^{EC}(o_3)$, they should utilize $S^{EC}(o_2)$ because they think the pressure comes from citizens with o_2. $S^{EC}(o_2)$ implies that obfuscators with o_3 should choose a_2 to maximize the entropy.

The opinion dynamics of the citizens can be summarized as follows:

- Citizens with o_2: their perception of the action-opinion relation is encoded in $S^{EC}(o_2)$. Non-obfuscators choose between a_2 and a_3 with equal probability according to $S^{EC}(o_2)$; obfuscators choose a_3 according to $S^{EC}(o_1|o_2)$. For both non-obfuscators and obfuscators, observing someone choosing a_3 makes them more confident in their current opinion. The inferring process after observing other actions (i.e., stage 1) relies on $S^{EC}(o_2)$ as described in Section 3.

- Citizens with o_3: their perception of the action-opinion relation is encoded in $S^{EC}(o_3)$. Non-obfuscators with o_3 will choose between a_1 and a_2 with equal probability according to $S^{EC}(o_3)$; obfuscators with o_3 will choose a_2 according to $S^{EC}(o_2)$. The *observing actions and inferring opinions* process (i.e., stage 1) relies on $S^{EC}(o_3)$ as described in Section 3.

Readers must have realized that this case seems to be more complex than what we presented in Section 3. This is because the assumption that agents have the

[5] Other forms of $S^{EC}(o_1|o_2)$ may also be feasible. We employ the current form because it helps illustrate the idea that different people have different perceptions of action-opinion relations, and they obfuscate based on different action-opinion matrices.

[6] In this example, $S^{EC}(o_1|o_2)$ and $S^{EC}(o_3)$ are composed of subsets of identical rows of $S^{EC}(o_2)$. This does not mean the rows in different matrices for the same opinion are always the same. They only depend on agent's perceptions. We thank the referee for pointing it out.

same action-opinion matrix has been relaxed. As we have discussed above, citizens with different opinions now have different perceptions of the action-opinion relation in this system. This is because citizens with o_2 have an imaginary type of neighbors: citizens with o_1. In addition, due to the different sources of social pressure (i.e., the motivation for hiding one's opinion through obfuscation), citizens also rely on different action-opinion matrices to decide how to obfuscate. Namely, obfuscators with o_2 believe that the pressure comes from citizens with o_1, therefore they choose actions according to the perception of these imaginary neighbors $S^{EC}(o_1|o_2)$. Meanwhile, obfuscators with o_3 believe that the pressure comes from citizens with o_2, therefore they rely on $S^{EC}(o_2)$ to hide opinions.

Indeed, the assumption that everyone knows and uses the same action-opinion matrix significantly simplifies the modeling process. Such a simplification is reasonable in many situations such as the vegetarian-omnivore case (Section 4.1), but here we show that it can be relaxed in order to capture the special mind-sets of different types of citizens.

Under a set of reasonable parameters and conditions, including (1) total population $N = 100$, (2) initial population of believers in o_2 and o_3 are equal, and (3) everyone knows everyone else in the system (i.e., G is a complete graph), we obtain the simulation results shown in Figure 4 (dynamics of opinions) and Figure 5 (dynamics of actions). If none of the citizens obfuscates (i.e., $N_o = 0$), o_2 and o_3 will dominate the population in turn, and the average population believing in each opinion over time is half of the whole population (Figure 4(a)). Meanwhile, about half of the population will keep silent (a_2), and the rest of the population is, on average, equally divided between citizens who mock the Emperor (a_1) and admire the Emperor (a_3) (Figure 5(a)).

To conclude, if no one wants to obfuscate, there is still a considerable number of citizens mocking the Emperor even when the majority is silent. However, as the number of obfuscators increases, the popularity of o_2 gradually grows (Figure 4(b), (c)). When everyone becomes an obfuscator ($N_o = 100$), it only takes a few time steps (note that each time step contains 100 individual updates) for the whole population to reach a consensus of o_2 (Figure 4(d)), that is, everyone becomes the "true believer" in Centola's model, and the "public lie" becomes the "(false) public opinion" that the Emperor is dressed. In terms of actions, everyone will eventually admire the Emperor when everyone obfuscates (Figure 5(d)).

Now let's consider the other extreme: what if citizens, opposite to obfuscation, would like to be as transparent as possible to observers? In other words, what if citizens want their opinions to be correctly and clearly known by others? Transparent citizens with o_2 will choose a_2 according to $S^{EC}(o_2)$: although a_3 has a smaller entropy, it is misleading because it signals that the underlying opinion is more likely to be o_1. They rely on their own perception

$S^{EC}(o_2)$ instead of $S^{EC}(o_1|o_2)$ (as the obfuscators do) because transparency is usually not directly related to the pressures from others. Meanwhile, transparent citizens with o_3, according to $S^{EC}(o_3)$, will choose a_1 because it directly signals that the underlying opinion is o_3. A population full of transparent citizens, with the same parameters and initial conditions as in Figure 4 and Figure 5, would produce a completely different result (Figure 6): in a few time steps, everyone will believe that the Emperor is naked (o_3), and mock the Emperor (a_1). In the context of Centola's model, this means everyone is now a "disbeliever", and the "public lie" is replaced by the "public truth".

To conclude, by applying the obfuscation-based opinion dynamics model we have provided an alternative explanation for the collective behavior in the tale by modeling obfuscation in public opinion formation. The phenomenon that everyone sincerely admires the invisible clothes can emerge from a population full of obfuscators. The fundamental difference with pluralistic ignorance is that in our analyses, citizens not only publicly admire the invisible clothes but also privately believe the clothes exist. On the contrary, if there are fewer obfuscators, eventually more citizens will believe that the Emperor is naked and dare to speak out the truth. Furthermore, if everyone would like to openly disclose their opinions (i.e., being transparent), there will soon be no believers in the swindlers' lie.

4.3. Qualitative conclusions from the examples

These two examples validate our early judgment that a universally correct answer to "how obfuscation affects public opinion" does not exist, but there are still some qualitative conclusions that worth mentioning. From the first example, we can arrive at a hypothesis that if an opinion only allows one action (vegetarianism in this example), it will be generally more popular than others in the presence of obfuscators. As argued in Section 4.1, this can be attributed to observer's overestimation of the popularity of this opinion. Although the hypothesis is not applicable in the second example (because o_1, the opinion that allows only one action, is imaginary), a similar logic can help us understand the simulation outcome. Obfuscators believing in o_2 play the same role as believers in o_1 because they always choose the same action a_3, therefore we could conceptually divide o_2 into two categories: o_2 that is believed by obfuscators (denoted by "obfuscating o_2") and o_2 that is believed by non-obfuscators (denoted by "non-obfuscating o_2"). Obfuscating o_2 can be viewed as an opinion that only allows one action, hence is expected to be more popular than other opinions (such as non-obfuscating o_2) according to the hypothesis. As a result, given the total number of believers in o_2 at any instance, the more people believe in obfuscating o_2, the more popular o_2 will be in the future. Meanwhile, the number of obfuscators (N_o) determines the initial number of believers in obfuscating o_2, and therefore is positively related to the popularity of o_2.

5. Conclusion and discussion

In the literature of opinion dynamics, we have witnessed two levels of details: most opinion dynamics models do not include the behavior of hiding opinions as they assume that opinions are always expressed publicly and truthfully; and studies into hiding opinions do not include the strategy of obfuscation as they assume that hiding opinions equates to keeping silent. These two omissions hamper our understanding of real-life opinion dynamics. This study contributes to the opinion dynamics literature by proposing an obfuscation-based opinion dynamics model that embodies a more complex and in some cases more realistic form of hiding opinions than keeping silent. The model embeds the obfuscation mechanism into the framework of the Action-Opinion Inference model, by formalizing a strategy of choosing the action that gives the least information about the underlying opinion.

For illustration purposes, we run the simulation of the model for two cases with different levels of complexity. The first vegetarian-omnivore case is relatively simple, providing a step-by-step demonstration. Simulation results indicate that in this particular case, obfuscation promotes the opinion (i.e., vegetarianism) that only allows one action while the more inclusive opinion (i.e., omnivorism) maintains a low popularity. The second case explains why the citizens in Han Christian Andersen's tale admire the Emperor's invisible clothes from the perspective of obfuscation. It is more complex because in this case obfuscators with different opinions have different perceptions of the relation between actions and opinions, and they rely on different perceptions to choose actions due to different motivations of obfuscation. The result suggests that obfuscation is able to facilitate the spread of the false opinion that the Emperor is dressed, while transparency can help popularize the true opinion that Emperor is naked.

Overall, the obfuscation-based opinion dynamics model expands the boundary of opinion dynamics studies by enabling agents to have a more intelligent strategy of hiding their opinions behind their actions. We hope that our study can initiate further discussions and developments about obfuscation and related notions. Directions of further research include (i) relaxing or modifying several assumptions such as undirected networks, positive influence, and sequential updating; (ii) calibrating the model to empirical data of public opinions to investigate obfuscation in real-world issues; and (iii) exploring concepts that are similar to (but subtly different from) obfuscation such as deception (Castelfranchi & Tan, 2001), strategic ambiguity (Eisenberg, 1984) and intentional vagueness (Blume & Board, 2014) as well as their roles in opinion dynamics.

Acknowledgments

This study was funded by the European Research Council as part of the Consolidator Grant BEHAVE (grant agreement No. 724431).

ORCID

Tanzhe Tang http://orcid.org/0000-0003-1504-9846
Amineh Ghorbani http://orcid.org/0000-0002-9985-8239
Caspar G. Chorus http://orcid.org/0000-0002-6380-4853

References

Acemoglu, D., Dahleh, M. A., Lobel, I., & Ozdaglar, A. (2011). Bayesian learning in social networks. *The Review of Economic Studies*, *78*(4), 1201–1236. https://doi.org/10.1093/restud/rdr004

Bjerring, J. C., Hansen, J. U., & Pedersen, N. J. L. L. (2014). On the rationality of pluralistic ignorance. *Synthese*, *191*(11), 2445–2470. https://doi.org/10.1007/s11229-014-0434-1

Blume, A., & Board, O. (2014). Intentional vagueness. *Erkenntnis*, *79*(4), 855–899. https://doi.org/10.1007/s10670-013-9468-x

Brunton, F., & Nissenbaum, H. (2015). *Obfuscation: A user's guide for privacy and protest*. The MIT Press.

Castelfranchi, C., & Tan, Y. H. (Eds.). (2001). *Trust and deception in virtual societies*. Kluwer.

Çelen, B., & Kariv, S. (2004). Observational learning under imperfect information. *Games and Economic Behavior*, *47*(1), 72–86. https://doi.org/10.1016/S0899-8256(03)00179-9

Centola, D., Willer, R., & Macy, M. (2005). The emperor's dilemma: A computational model of self-enforcing norms. *American Journal of Sociology*, *110*(4), 1009–1040. https://doi.org/10.1086/427321

Chorus, C., Van Cranenburgh, S., Daniel, A. M., Sandorf, E. D., Sobhani, A., & Szép, T. (2020). Obfuscation maximization-based decision-making: Theory, methodology and first empirical evidence. *Mathematical Social Sciences*, *109*, 28–44. https://doi.org/10.1016/j.mathsocsci.2020.10.002

Davis, R. C. (2019). Obfuscating authorship: Results of a user study on nondescript, a digital privacy tool [working paper]. CUNY Academic Works.

De Jaegher, K. (2003). A game-theoretic rationale for vagueness. *Linguistics and Philosophy*, *26*(5), 637–659. https://doi.org/10.1023/A:1025853728992

Dewan, T., & Myatt, D. P. (2008). The qualities of leadership: Direction, communication, and obfuscation. In *American political science review 102*(3)(pp. 351–368). doi:10.1017/S0003055408080234

Doyle, T. (2018). Privacy, obfuscation, and propertization. *IFLA Journal*, *44*(3), 229–239. https://doi.org/10.1177/0340035218778054

Eisenberg, E. M. (1984). Ambiguity as strategy in organizational communication. *Communication Monographs*, *51*(3), 227–242. https://doi.org/10.1080/03637758409390197

Fan, K., & Pedrycz, W. (2015). Emergence and spread of extremist opinions. *Physica A: Statistical Mechanics and Its Applications*, *436*, 87–97. https://doi.org/10.1016/j.physa.2015.05.056

Fan, K., & Pedrycz, W. (2016). Opinion evolution influenced by informed agents. *Physica A: Statistical Mechanics and Its Applications*, *462*, 431–441. https://doi.org/10.1016/j.physa.2016.06.110

Flache, A., Mäs, 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), 4. https://doi.org/10.18564/jasss.3521

Gale, D., & Kariv, S. (2003). Bayesian learning in social networks. *Games and Economic Behavior*, *45*(2), 329–346. https://doi.org/10.1016/S0899-8256(03)00144-1

Gawronski, P., Nawojczyk, M., & Kulakowski, K. (2014). Opinion formation in an open system and the spiral of silence. *arXiv Preprint arXiv:1407.2742*.

Golub, B., & Sadler, E. (2016). Learning in social networks. In A. Galeotti & B. W. Roger (Eds.), *The oxford handbook of the economics of networks* (pp. 504–542). Oxford University Press.

Golub, B., & Jackson, M. O. (2010). Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, *2*(1), 112–149. doi:10.1257/mic.2.1.112.

Grandi, U., Lorini, E., Novaro, A., & Perrussel, L. (2017). Strategic disclosure of opinions on a social network. In *Proceedings of the 16th conference on autonomous agents and multi agent systems* (pp. 1196–1204). Sao Paulo, Brazil.

Hansen, J. U. (2012). A logic-based approach to pluralistic ignorance. In J. D. Vuyst & L. Demey (Eds.), *Future directions for logic—proceedings of PhDs in Logic III* (pp. 67–80). College Publications.

Jadbabaie, A., Molavi, P., Sandroni, A., & Tahbaz-Salehi, A. (2012). Non-Bayesian social learning. *Games and Economic Behavior*, *76*(1), 210–225. https://doi.org/10.1016/j.geb.2012.06.001

Krapivsky, P. L., Redner, S., & Ben-Naim, E. (2010). *A kinetic view of statistical physics*. Cambridge University Press.

Lobel, I., Acemoglu, D., Dahleh, M., & Ozdaglar, A. (2009). Rate of convergence of learning in social networks. In *Proceedings of the American Control Conference* (pp. 2825–2830), St. Louis, MO, USA.

Miller, D. T., & McFarland, C. (1987). Pluralistic ignorance: When similarity is interpreted as dissimilarity. *Journal of Personality and Social Psychology*, *53*(2), 298. https://doi.org/10.1037/0022-3514.53.2.298

Mitsutsuji, K., & Yamakage, S. (2020). The dual attitudinal dynamics of public opinion: An agent-based reformulation of LF Richardson's war-moods model. *Quality & Quantity*, *54*(2), 439–461. https://doi.org/10.1007/s11135-019-00938-x

Mossel, E., Olsman, N., & Tamuz, O. (2016). Efficient bayesian learning in social networks with gaussian estimators. In *54th annual allerton conference on communication, control, and computing (allerton)* (pp. 425–432). IEEE, Monticello, IL, USA.

Noelle-Neumann, E. (1974). The spiral of silence a theory of public opinion. *Journal of Communication*, *24*(2), 43–51. https://doi.org/10.1111/j.1460-2466.1974.tb00367.x

Ross, B., Pilz, L., Cabrera, B., Brachten, F., Neubaum, G., & Stieglitz, S. (2019). Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks. *European Journal of Information Systems*, *28*(4), 394–412. https://doi.org/10.1080/0960085X.2018.1560920

Rossman, G. (2014). Obfuscatory relational work and disreputable exchange. *Sociological Theory*, *32*(1), 43–63. https://doi.org/10.1177/0735275114523418

Schilke, O., & Rossman, G. (2018). It's only wrong if it's transactional: Moral perceptions of obfuscated exchange. *American Sociological Review*, *83*(6), 1079–1107. https://doi.org/10.1177/0003122418806284

Serra-Garcia, M., Van Damme, E., & Potters, J. (2011). Hiding an inconvenient truth: Lies and vagueness. *Games and Economic Behavior*, *73*(1), 244–261. https://doi.org/10.1016/j.geb.2011.01.007

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Sohn, D. (2019). Spiral of silence in the social media era: A simulation approach to the interplay between social networks and mass media. In *Communication Research* . https://doi.org/10.1177/0093650219856510

Sohn, D., & Geidner, N. (2016). Collective dynamics of the spiral of silence: The role of ego-network size. *International Journal of Public Opinion Research*, *28*(1), 25–45. https://doi.org/10.1093/ijpor/edv005

Takeuchi, D., Tanaka, G., Fujie, R., & Suzuki, H. (2015). Public opinion formation with the spiral of silence on complex social networks. *Nonlinear Theory and Its Applications*, IEICE, *6*(1), 15–25. https://doi.org/10.1587/nolta.6.15.

Tang, T., & Chorus, C. G. (2019). Learning opinions by observing actions: Simulation of opinion dynamics using an action-opinion inference model. *Journal of Artificial Societies and Social Simulation*, *22*(3), 3. https://doi.org/10.18564/jasss.4020

Wherry, F. F., Seefeldt, K. S., & Alvarez, A. S. (2019). To lend or not to lend to friends and kin: Awkwardness, obfuscation, and negative reciprocity. *Social Forces*, *98*(2), 753–793. doi: 10.1093/sf/soy127

You, I., & Yim, K. (2010). Malware obfuscation techniques: A brief survey. In *2010 international conference on broadband, wireless computing, Communication and Applications*, 297–300. Fukuoka, Japan

Appendix

In this Appendix, we derive an analytical solution of the dynamics described in Section 4.1, namely the Vegetarian-Omnivore example, whose simulation results have already been given in Figure 2 and Figure 3. The derivation closely follows Tang and Chorus[7] (2019), which itself is an extension of a typical derivation of the voter model (Krapivsky et al., 2010). It should be noted that the derivation significantly benefits from the simplicity of S^{VO}, and is therefore only applicable to this particular case.

Recall the action-opinion matrix used in the example:

$$S^{VO} = \begin{array}{ccc} & \mathit{Steak} & \mathit{Salad} \\ \mathit{Veg} & -1 & +1 \\ \mathit{Omn} & 0 & 0 \end{array}$$

We start by rephrasing the notion of opinions in a binary fashion: denote the opinion of an agent i as a binary variable $o^{(i)}$ which can only take one of two values ± 1. $o^{(i)} = 1$ means the agent is a vegetarian, and $o^{(i)} = -1$ means the agent is an omnivore. Additionally, we denote $a_1 = \mathit{Steak}$, and $a_2 = \mathit{Salad}$.

The probability that agent i changes her opinion ("flip rate"), based on Eq. (2) and (3), can be written as:

$$w_i = \frac{1}{2}\left\{1 - \frac{o^{(i)}}{z}\left[\sum_{j \in M_i} \sum_{o \in \{\pm 1\}} P\left(o^{(j)} = o | a^{(j)}\right) o\right]\right\} \tag{A1}$$

where z is the number of neighbors of each agent (i.e., lattice coordination number) and is assumed to be constant. M_i is the set of all the neighbors of agent i. $P\left(o^{(i)} = o|a^{(j)}\right)$ is equivalent to $P^{(i)}\left(o^{(i)} = o|a^{(j)}\right)$ as the inference is the same for everyone who observes $a^{(j)}$.

Following Tang and Chorus (2019), we focus on the average opinion of each agent $R(i,t) \equiv < o^{(i)}(t) >$. Note that by "$< \cdot >$" we mean the average $< F(X) > \equiv \sum_x P(X = x)F(x)$, therefore $R(i,t) \equiv \sum_o P\left(o^{(i)}(t) = o\right) o$ is the opinion of agent i averaged over all possible values of $o^{(i)}$, which can be roughly interpreted as the opinion of agent i averaged over all (countless) realizations (or "trajectories" in the language of statistical physics) of the dynamics at time t. It is neither the average opinion of all agents nor agent i's opinion averaged over time. To be precise, suppose there are Q systems (i.e., realizations) $S_1, \ldots, S_q, \ldots, S_Q$ that all evolve independently from the same initial system S_0, and the opinion of agent i in each system S_q at time t is $o_q^{(i)}(t)$, then $R(i,t) = \lim_{Q \to \infty} \frac{\sum_{q=1}^{Q} o_q^{(i)}(t)}{Q}$. To keep things tidy, we omit t and write $R(i)$.

The paper describes a discrete-time model where an agent is chosen to update her opinion in a time unit, and N successive time units define a time step. The discreteness helps implement simulation but not derivation. Here, we alternatively assume that time (t) in the dynamics is continuous to facilitate the derivation. The continuous alternative, as we will witness at the end of this Appendix, can produce good approximation of the discrete model given sufficiently long time.

In a continuous-time context, the dynamics of agent i's opinion in a sufficiently short time interval Δt is:

[7] Most part of the derivation was modified from Tang and Chorus (2019).

$$o^{(i)}(t+\Delta t)=\begin{cases} o^{(i)}(t) & \text{with probability } 1-w_i\Delta t \\ -o^{(i)}(t) & \text{with probability } w_i\Delta t \end{cases} \tag{A2}$$

According to Krapivsky et al. (2010), the evolution of $R(i)$ is:

$$\frac{dR(i)}{dt}=\frac{d<o^{(i)}>}{dt}=-2\left\langle o^{(i)}w_i\right\rangle \tag{A3}$$

The derivation of the last term is based on (A2). By substituting (A1) into (A3) and using the trick that $\left[o^{(i)}\right]^2=1$, we obtain:

$$\frac{dR(i)}{dt}=-R(i)+\frac{1}{z}\sum_{j\in M_i}<\sum_o P\left(o^{(j)}=o|a^{(j)}\right)o> \tag{A4}$$

By denoting $<\sum\limits_o P\left(o^{(j)}=o|a^{(j)}\right)o>$ as $R^*(j)$, (A4) can be expressed in a more elegant form:

$$\frac{dR(i)}{dt}=-R(i)+\frac{1}{z}\sum_{j\in M_i}R^*(j) \tag{A5}$$

To describe the whole population, we define the "mean magnetization" (analogous to the same concept in spin dynamics) $m\equiv\frac{1}{N}\sum_i R(i)$, which is the average opinion of the population averaged over all realizations. The mean magnetization m represents public opinion: if $m=1$, everyone is a vegetarian in all realizations without exception; if $m=-1$, everyone is an omnivore in all realizations without exception; if $m=0$, the population as a whole does not have a preference. Note that $\frac{dm}{dt}=\frac{d\left(\frac{1}{N}\sum_i R(i)\right)}{dt}=\frac{1}{N}\sum\limits_i\frac{dR(i)}{dt}$, then summing (A5) over all agents leads to:

$$N\frac{dm}{dt}=-\sum_i R(i)+\frac{1}{z}\sum_i\sum_{j\in M_i}R^*(j) \tag{A6}$$

Note that $R(i)\equiv\left\langle o^{(i)}\right\rangle=\sum\limits_o P\left(o^{(i)}=o\right)o$ and o can only take two values ± 1, we have:

$$R(i)=P\left(o^{(i)}=1\right)-P\left(o^{(i)}=-1\right)=2P\left(o^{(i)}=1\right)-1 \tag{A7}$$

Similarly, we have $R^*(j)=\left\langle P\left(o^{(j)}=1|a^{(j)}\right)-P\left(o^{(j)}=-1|a^{(j)}\right)\right\rangle$, hence:

$$R^*(j)=2\left\langle P\left(o^{(j)}=1|a^{(j)}\right)\right\rangle-1 \tag{A8}$$

According to the definition of "$\langle\cdot\rangle$":

$$\begin{aligned}\left\langle P\left(o^{(j)}=1|a^{(j)}\right)\right\rangle &= P\left(o^{(j)}=1|a^{(j)}=a_1\right)P\left(a^{(j)}=a_1\right)\\ &\quad+P\left(o^{(j)}=1|a^{(j)}=a_2\right)P\left(a^{(j)}=a_2\right)\end{aligned} \tag{A9}$$

From S^{VO}, we know that $P\left(o^{(j)}=1|a^{(j)}=a_1\right)=0$, and $P\left(o^{(j)}=1|a^{(j)}=a_2\right)=\frac{2}{3}$. Substituting them into (A9), we have:

$$\left\langle P\left(o^{(j)}=1|a^{(j)}\right)\right\rangle=\frac{2}{3}P\left(a^{(j)}=a_2\right) \tag{A10}$$

Substituting (A7), (A8), and (A10) into (A6):

$$N\frac{dm}{dt} = -2\sum_i P\left(o^{(i)} = 1\right) + \frac{4}{3z}\sum_i \sum_{j\in M_i} P\left(a^{(j)} = a_2\right) \tag{A11}$$

Note that:

$$\sum_i \sum_{j\in M_i} P\left(a^{(j)} = a_2\right) = z\sum_i P\left(a^{(i)} = a_2\right) \tag{A12}$$

because everyone has been counted z times. Substituting (A12) into (A11):

$$N\frac{dm}{dt} = 2\sum_i \left(\frac{2}{3}P\left(a^{(i)} = a_2\right) - P\left(o^{(i)} = 1\right)\right) \tag{A13}$$

From (A13) we know that z has been canceled out. This means the number of neighbors of each agent does not affect the dynamics of m, as long as everyone has the same number of neighbors.

Until (A13), what we have done is simply modifying the derivation of the AOI model by Tang and Chorus (2019). From now on, we start to take into account obfuscation. Suppose the share of obfuscators in the population is θ $(0 \le \theta \le 1)$. In addition, we introduce another binary variable $ob^{(i)}$: $ob^{(i)} = 1$ means agent i is an obfuscator, and $ob^{(i)} = -1$ means agent i is not an obfuscator. Using this notion, we have:

$$\begin{cases} P\left(a^{(i)} = a_2 | o^{(i)} = 1\right) = 1 \\ P\left(a^{(i)} = a_2 | o^{(i)} = -1, ob^{(i)} = 1\right) = 1 \\ P\left(a^{(i)} = a_2 | o^{(i)} = -1, ob^{(i)} = -1\right) = 0.5 \end{cases} \tag{A14}$$

and

$$\begin{cases} P\left(o^{(i)} = -1, ob^{(i)} = 1\right) = P\left(o^{(i)} = -1\right)P\left(ob^{(i)} = 1\right) = \theta P\left(o^{(i)} = -1\right) \\ P\left(o^{(i)} = -1, ob^{(i)} = -1\right) = P\left(o^{(i)} = -1\right)P\left(ob^{(i)} = -1\right) = (1-\theta)P\left(o^{(i)} = -1\right) \end{cases} \tag{A15}$$

The derivation of (A15) is based on the fact that being an obfuscator or not is independent of one's opinion. Meanwhile, we can expand $P\left(a^{(i)} = a_2\right)$:

$$\begin{aligned} P\left(a^{(i)} = a_2\right) = {} & P\left(o^{(i)} = 1\right)P\left(a_2 | o^{(i)} = 1\right) \\ & + P\left(o^{(i)} = -1, ob^{(i)} = 1\right)P\left(a_2 | o^{(i)} = -1, ob^{(i)} = 1\right) \\ & + P\left(o^{(i)} = -1, ob^{(i)} = -1\right)P\left(a_2 | o^{(i)} = -1, ob^{(i)} = -1\right) \end{aligned} \tag{A16}$$

Substituting (A14) and (A15) into (A16):

$$P\left(a^{(i)} = a_2\right) = \frac{1}{2}\left[(1-\theta)P\left(o^{(i)} = 1\right) + (1+\theta)\right] \tag{A17}$$

By substituting (A17) into (A13), we obtain:

$$N\frac{dm}{dt} = 2\sum_i \left\{\frac{1}{3}\left[(1-\theta)P\left(o^{(i)} = 1\right) + (1+\theta)\right] - P\left(o^{(i)} = 1\right)\right\} \tag{A18}$$

From (A7) we know $R(i) = 2P\left(o^{(i)} = 1\right) - 1$, therefore:

$$m \equiv \frac{1}{N}\sum_i R(i) = \frac{2}{N}\sum_i P\left(o^{(i)} = 1\right) - 1 \tag{A19}$$

According to (A19), (A18) can be rewritten as:

$$\frac{dm}{dt} = -\frac{\theta + 2}{3}m + \frac{\theta}{3} \tag{A20}$$

The stable fixed point of (A20) is:

$$m = \frac{\theta}{2 + \theta} \tag{A21}$$

which is equivalent to:

$$\frac{1}{N}\sum_i P\left(o^{(i)} = 1\right) = \frac{1 + \theta}{2 + \theta} \tag{A22}$$

(A22) tells us that at equilibrium, the share of vegetarians in the population is $\frac{1+\theta}{2+\theta}$. However, this does not mean for every realization, the system will converge to this equilibrium. Instead, the share of vegetarians averaged over all realizations of the dynamics will converge to $\frac{1+\theta}{2+\theta}$. This result is in line with Figure 3, where $\bar{f}_{Veg}$ averaged over all the realizations carried out in the simulation (which is not "all realizations" but an approximation of "all realizations") is well approximated by $\frac{1+\theta}{2+\theta}$.