Modeling Opinion-Behavior Co-evolution for Explaining Opinion Polarization: A Framework

Tanzhe Tang, Amineh Ghorbani, Caspar G. Chorus

THESIS CHAPTER

Chapter 5 (pp. 111-154) of the doctoral dissertation *On the Methods for Explaining Polarization of Private and Unobservable Opinions: An Opinion-Behavior Co-Evolutionary Approach*, Delft University of Technology, 2022. This standalone file reproduces the final published chapter and adds only this bibliographic cover page.

ABSTRACT

The dynamics of opinions and behaviors are almost inseparable. Therefore, modeling opinion-behavior co-evolution is expected to be a more realistic approach to explaining opinion polarization than modeling opinion dynamics alone. Regardless of the co-evolution models that have already been proposed in various disciplines, this novel approach is still largely uncoordinated, and a unifying framework is in urgent need to organize existing efforts and facilitate future studies. In this paper, such a framework, called MOBEP (Modeling Opinion-Behavior Co-evolution for Explaining Opinion Polarization), is developed based on the central notion that behaviors serve as the messenger of opinions, highlighting the inference process that translates observed behaviors into inferred opinions. For validation purposes, the framework is applied to a selection of representative models. Finally, a case study about mask wearing during the COVID-19 pandemic is presented as a vivid demonstration of how the framework works.

RECOMMENDED CITATION

Tang, T., Ghorbani, A., & Chorus, C. G. (2022). Modeling opinion-behavior co-evolution for explaining opinion polarization: A framework. In T. Tang, *On the Methods for Explaining Polarization of Private and Unobservable Opinions: An Opinion-Behavior Co-Evolutionary Approach* (Chapter 5, pp. 111-154). Delft University of Technology.

SOURCE

Final published dissertation:
https://doi.org/10.4233/uuid:b4b02f3b-b622-4226-8f8c-be26e2c52428
License: CC BY

5

MODELING OPINION-BEHAVIOR CO-EVOLUTION FOR EXPLAINING OPINION POLARIZATION: A FRAMEWORK

*The dynamics of opinions and behaviors are almost inseparable. Therefore, modeling opinion-behavior co-evolution is expected to be a more realistic approach to explaining opinion polarization than modeling opinion dynamics alone. Regardless of the co-evolution models that have already been proposed in various disciplines, this novel approach is still largely uncoordinated, and a unifying framework is in urgent need to organize existing efforts and facilitate future studies. In this paper, such a framework, called MOBEP (**M**odeling **O**pinion-**B**ehavior Co-evolution for **E**xplaining Opinion **P**olarization), is developed based on the central notion that behaviors serve as the messenger of opinions, highlighting the inference process that translates observed behaviors into inferred opinions. For validation purposes, the framework is applied to a selection of representative models. Finally, a case study about mask wearing during the COVID-19 pandemic is presented as a vivid demonstration of how the framework works.*

Keywords: Opinion-behavior co-evolution, Opinion polarization, Agent-based modeling framework

5.1. INTRODUCTION

Opinion polarization can be roughly viewed as the degree to which a population can be separated into two antagonistic groups in terms of their opinions. From the fundamental disagreement between pro- and anti-vaccination views to the sharp cleavage between Republicans and Democrats, opinion polarization is not only omnipresent but also closely

related to conflicts. Hence, finding mechanisms to explain opinion polarization (by mathematical/ computational modeling) has become one of the most popular and promising topics in the field of opinion dynamics, while recent worldwide crises such as the global pandemic, climate change, and political unrest are fueling both academic and public interest in this topic to an even greater extent.

To explain opinion polarization, several strands of opinion dynamics models have been proposed, including models of rejection (Macy et al., 2003; Flache & Macy, 2011), bounded confidence (Deffuant et al., 2000; Hegselmann & Krause, 2002), persuasive arguments (Mäs & Flache, 2013; Mäs et al., 2013), and social feedbacks (Banisch & Olbrich, 2019). An intuitively important but often ignored factor in explaining opinion polarization is people's behavior. Opinions and behaviors are a strongly interacting pair that often come hand-in-hand. In reality, it is usually very difficult to disentangle one from the other. For example, we learn opinions of others from their behaviors, while our own behaviors are (at least partly) governed by our opinions. As one's opinion is private and unobservable, opinion polarization models that exclude behavior are in fact assuming a direct communication manner that people always express their opinions truthfully, and other people's interpretation is always correct. Such an assumption, as identified by many scholars (Tang & Chorus, 2019; Mitsutsuji & Yamakage, 2020; Zhan et al., 2021), is at least not always realistic because it ignores many possibilities such as deception, obfuscation, and misinterpretation. Hence, a mechanism that takes into account behavior is expected to deliver a more realistic explanation of opinion polarization.

Even from a purely theoretical perspective, behavior still deserves more attention as incorporating behavior in opinion dynamics opens up the possibility of developing and exploring various new mechanisms of opinion polarization. In fact, there are already a few polarization mechanisms where behavior plays a central role. A remarkable example is the CODA model (Martins, 2008) where people have binary behaviors and continuous opinions. The basic idea is that an individual's opinion is intensified after observing others doing the same behavior as hers. Martins (2008) shows that when applied to the voter model (Clifford & Sudbury, 1973; Holley & Liggett, 1975) and the Sznajd model (Sznajd-Weron & Sznajd, 2000; Stauffer, 2003; Sznajd-Weron, 2005), the mechanism is able to produce a bimodal distribution of opinions, which is a clear indicator of polarization. Meanwhile, the AOI model (Tang & Chorus, 2019) uses a set of predefined action-opinion relations to guide the interactions between discrete behaviors and discrete opinions, and it has been shown that under certain types of action-opinion relations, a polarized opinion distribution is likely to emerge. In addition, as we will show in Section 5.3, the social feedback model (Banisch & Olbrich, 2019) can also be viewed as an example of opinion polarization models that include both opinion and behavior. Regardless of their different theoretical assumptions and model settings, these examples together exhibit the potential of (modeling) behavior as a key element in explaining opinion polarization.

In addition, a number of models and mechanisms that describe the co-evolution of opinions and behaviors have been found in various disciplines (e.g., Gawronski et al., 2014; Huang & Wen, 2014; Buechel et al., 2015; Grandi et al., 2017; Mitsutsuji & Yamakage, 2020; Zino et al., 2020a, b; Zhan et al., 2021). Although not being designed to explain polarization, they still offer valuable insights into modeling the co-evolution in these different respects, which will be summarized in Section 5.2. However, due to

barriers between disciplines and differences between terminologies, these works are largely uncoordinated with limited acknowledgment of each other, and their insights, in general, are less well known in the opinion polarization community.

To organize existing efforts in modeling opinion-behavior co-evolution and facilitate future studies of opinion polarization, we propose a unifying framework, called MOBEP (**M**odeling **O**pinion-**B**ehavior Co-Evolution for **E**xplaining Opinion **P**olarization) framework, that encompasses various mechanisms and models including those discussed above. The framework aims to provide a starting point for computational sociologists – especially agent-based modelers – to conceptualize and analyze opinion-behavior co-evolution, and thereby advance model development and knowledge accumulation in opinion polarization. As a demonstration of the framework, we provide a case study of mask wearing during the COVID-19 pandemic, where opinions (preferences for mask wearing) and behaviors (wear a mask or not) are closely intertwined.

The rest of the paper is organized as follows. Section 5.2 provides a systematic review of models of behavior-opinion co-evolution. The MOBEP framework will be introduced Section 5.3, followed by the case study in Section 5.4. Section 5.5 presents summary and discussion, together with outlooks to further research.

5.2. BACKGROUND

Classic opinion dynamics models, including models of opinion polarization, routinely assume that opinions are observable and that one's opinion can directly affect others' opinions. In this section, we will review models that deviate from this assumption by including behavior in the dynamics. Note that different models may use different names for "opinion" and "behavior", but as we will soon find out, they are in nature describing the same type of dynamics. In general, these models can be categorized into two groups according to the type of behaviors in the co-evolution, namely the behavior of expressing one's opinion and behavior in a general sense.

5.2.1. MODELS OF OPINION EXPRESSION

"Expressing one's opinion" should be the behavior that is most related to opinions, and it is not surprising to find out that this behavior has appeared in many opinion dynamics models. These models can be roughly divided into "dual opinion models" whose focus is the relations between expressed opinion and private opinion, and "silence models" where an agent decides whether to express her opinion or not.

DUAL OPINION MODELS

The difference between one's privately held and publicly expressed opinions has motivated a number of opinion dynamics models (e.g., Huang & Wen, 2014; Buechel et al., 2015; Gastner et al., 2018; Shang, 2019; Ye et al., 2019; Mitsutsuji & Yamakage, 2020), where an agent's expressed opinion is usually a joint result of both what she really believes (i.e., her private held opinion) and what others have expressed (i.e., other agents' expressed opinions). For example, in Huang & Wen's (2014) model of pluralistic ignorance, there are three possible outcomes when an agent faces others' expressed opinions: (i) both her private and expressed opinions stay unchanged; (ii) both her private and expressed opinions are affected by others' expressed opinion (called "private acceptance"); and

(iii) only her expressed opinion is influenced while her private opinion stays unchanged (called "public compliance"). The outcome for a particular agent is determined by many factors, including the agent's level of uncertainty and the pressure to conform. Buechel et al. (2015)'s model of conformity assumes that one's expressed opinion is motivated by both honesty (the desire to be in line with one's own private opinion) and conformity (the desire to align with the majority's expressed opinion). The assumption is implemented by a utility function that takes into account both the difference between the agent's expressed opinion and her private opinion, and the difference between her expressed opinion and the average expressed opinion of others. Recently, Mitsutsuji & Yamakage (2020) develop a discrete opinion model for war-moods, where they make the distinction between the "private sphere" and the "public sphere". In the private sphere, agents freely exchange their private opinions, while in the public sphere, only publicly expressed opinions are observed, and agents may adopt an expressed opinion that is different from their private opinions due to the pressure to conform. Formally, in the public sphere, an agent chooses opinion θ with the probability $P(\theta) \propto r(\theta)^x$, where $r(\theta)$ is the popularity of θ in the agent's neighborhood, and x is a parameter accounting for the pressure to conform[1].

SILENCE MODELS

Unlike dual opinion models where the task for agents is to decide how to publicly express their opinions, agents in what we call silence models need to decide whether they should express their opinions or not. Many silence models are based on the spiral of silence theory, postulating that people's willingness to express their private opinions decreases if they feel that their opinions are unpopular (Noelle-Neumann, 1974). Accordingly, spiral of silence models have developed various mechanisms for agents to switch between "speaking out" and "keeping silent". For example, Gawronski et al. (2014) assume that an agent is more likely to express her private opinion if it is not too different from the average expressed opinion of others weighted by their charismas. In the model by Sohn and Geidner (2016), the condition of expressing one's opinion is that the intensity of her private opinion (in the form of the opinion's absolute value) exceeds a given threshold. Similarly, Ross et al. (2019) assume that an agent will express her opinion if her confidence, which is determined by the observed opinions expressed by others, exceeds her willingness to self-censor, and will keep silent otherwise. In Gaisbauer et al. (2020)'s model, besides comparing the expected population of publicly agreeing and disagreeing neighbors, agents also need to consider the "cost of opinion expression" when they decide whether to speak out or not. In addition to models of spiral of silence, Grandi et al. (2017) and later Shepherd & Goldsmith (2020) have discussed the topic of "strategic disclosure of opinions" (Grandi et al., 2017), meaning that agents strategically choose to express opinions or not in order to achieve their individual goals.

To summarize, models of opinion expression describe a particular type of opinion-behavior co-evolution where the behavior is simply "expressing an opinion". They extend the boundary of classic opinion dynamics models by deviating from the implicit assumption that opinions are always truthfully expressed. However, the boundary still needs to be further extended if we would like to model behavior in a more generic manner.

[1] In the original model, an agent still has a small probability to express her private opinion regardless of what others are expressing. In addition, the model also includes exogenous impacts such as the outbreak of war, but they are out of our scope.

5.2.2. MODELS OF GENERAL BEHAVIOR

Opinions can direct numerous behaviors aside from expressing opinions. Here we briefly introduce some representative models that describe the co-evolution of opinions and general behaviors. As we will see soon, when the relation between opinion and behavior is sufficiently simple (e.g., CODA model, CMAO, and SNOAEs), the general behavior is in essence the same as "expressing opinions".

CODA MODEL

The most well-known model that brings the concept of behavior to opinion polarization studies is probably the Continuous Opinion Discrete Action (CODA) model (Martins, 2008) (as "behavior" and "action" are of no difference in our context). In the CODA model, behavior is a binary choice that can be observed by others, while opinion is the unobservable probability of an agent that one behavior is better than the other. Denote the behavior of agent i as $m_i \in \{-1, 1\}$ where -1 and 1 are the two options, and the probability that agent i thinks 1 is better than -1 as p_i. The relation between opinion and behavior is then $m_i = sign(1 - p_i)$ (Martins, 2008, 2014). An agent i believes that her neighbor j will choose $m_j = 1$ (-1) if 1 (-1) is the better option with a fixed probability $\alpha > 0.5$. Through simple calculations using the Bayes theorem, Martins finds that if agent i observes a neighbor j choosing 1 (i.e., $m_j = 1$), her "log-odds transform" of opinion $v_i = ln(p_i/(1-p_i))$ will be updated to $v_i + ln(\alpha/(1-\alpha))$, and if neighbor j chooses -1 (i.e., $m_j = -1$), v_i will be updated to $v_i - ln(\alpha/(1-\alpha))$ (Martins, 2008, 2014). In the original paper (Martins, 2008), the CODA model is implemented according to interaction rules of the voter model (Clifford & Sudbury, 1973; Holley & Liggett, 1975) and the Sznajd model (Sznajd-Weron & Sznajd, 2000; Stauffer, 2003; Sznajd-Weron, 2005) respectively, and polarized opinion distributions have been observed to emerge in both implementations. The reason why the CODA model can produce polarization is probably related to the assumption that opinions are continuous but actions (behaviors) are discrete (in fact, binary). The discrete behaviors serve as a "classifier" of the continuous opinions in the sense that all agents with an opinion larger (smaller) than 0.5 will choose the same behavior. As a result, an agent with a moderate opinion (e.g., $o_i = 0.51$) and an agent with an extreme opinion of the same direction (e.g., $o_i = 1$) exert the same influence on their observers as they will both choose the behavior $m_i = 1$. For example, when an agent with an opinion of 0.6 observes the behavior of her neighbor whose opinion is 0.51, the agent's opinion will become closer to 1 although her neighbor is actually less extreme than herself.

AOI MODEL

Unlike the CODA model, the Action-Opinion Inference (AOI) model (Tang & Chorus, 2019) assumes that both opinions and behaviors are discrete, which makes it possible to use a simple deontic logic to represent the relations between opinion and behavior: a behavior may be obliged, permitted, or prohibited by an opinion. An agent must choose a behavior if it is obliged by her opinion, but if her opinion permits more than one behavior, she then chooses one of them randomly. Intuitively, the forbidden behavior will never be chosen. Knowing the relations between opinions and behaviors, an agent can infer the opinions of her neighbors from their behaviors in a Bayesian manner, even if opinions are not observable. The inferred opinions then influence the agent's own opinion as in a

classic opinion dynamics model. For example, suppose there are two competing opinions about diet: vegetarianism and omnivorism, and there are only two meals available: beef and salad. Eating beef is forbidden by vegetarianism but permitted (not obliged) by omnivorism, while eating salad is obliged by vegetarianism and permitted by omnivorism. This relation determines that a vegetarian agent will never eat beef, but an omnivore, without any additional information about her preference, would choose between beef and salad with equal probability. By observing another agent's behavior, an agent infers the underlying opinion through the "action-opinion inference process", which is a reverse of how behaviors are determined according to opinions. In the example above, seeing someone eating beef directly reveals that she is an omnivore, while eating salad is a less informative signal as she can be either a vegetarian or an omnivore. With these inferred opinions, agents update their own opinions and, in the next time step, update their behaviors accordingly. The highlight of the AOI model is the introduction of the deontic behavior-opinion relations that can take various forms, based on which the model can produce consensus, polarization[2], and diversity.

CO-EVOLUTIONARY MODEL OF ACTIONS AND OPINIONS

Zino et al. (2020a, b) propose the Co-evolutionary Model of Actions and Opinions (hereafter abbreviated as "CMAO") by combining ideas from both opinion dynamics and game theory. Apart from modeling how agents' opinions are affected by the behaviors of others (called "influence layer"), CMAO also includes the direct influence between opinions (called "communication layer"). In the influence layer, agents can only observe each other's behaviors but not their opinions, while in the communication layer, agents can directly observe each other's opinions. The two layers are indeed similar to what Mitsutsuji and Yamakage (2020) called "public sphere" and "private sphere". In short, agents in CMAO simultaneously update their (binary) behaviors and (continuous) opinions. An agent's behavior is determined by both her own opinion and the observed behaviors of others in the influence layer via a decision-making mechanism, and her opinion is updated according to the shared opinions in the communication layer and the behaviors observed from the influence layer via an opinion dynamics mechanism. CMAO integrates the ideas of classic opinion dynamics models that people's opinions are directly influenced by others' opinions, and the ideas of the CODA and AOI model that opinions are learned by observing behaviors. To cite the authors, opinion dynamics and behavior dynamics "are coupled seamlessly, while each separate dynamics inherits the fundamental features of their separate grounding frameworks" (Zino et al., 2020b).

SOCIAL NETWORK OPINIONS AND ACTIONS EVOLUTIONS MODEL

The Social network opinions and actions evolutions (SNOAEs) model by Zhan et al. (2021) is fundamentally similar to CMAO. In SNOAEs, an agent's binary behavior is determined solely by her continuous opinion: if her opinion is larger than the opinion threshold, she will choose one behavior, and she will choose another behavior otherwise. The social network in SNOAEs plays a central role in opinion dynamics. If a randomly chosen pair

[2] Note that in the AOI model, opinions do not have values so that we cannot measure the similarities between opinions. As a result, a polarized opinion distribution is defined as the co-existence of a limited number of opinions with similar numbers of believers.

of agents are connected via the network, the focal agent can directly know the opinion of her partner and update her own opinion using the bounded confidence mechanism (Deffuant et al., 2000; Hegselmann & Krause, 2002). If they are not connected, the focal agent can only observe her partner's behavior. The observed behavior is then used as her partner's opinion to update the focal agent's opinion in the same manner as in the case where they are connected.

By ignoring the details of implementation, we can see that what happens when the pair of interacting agents are connected is the same as the dynamics in the "private sphere" of Mitsutsuji & Yamakage (2020), or the "communication layer" of Zino et al. (2020a, b), and what happens when they are not connected is the same as the dynamics in the "public sphere" (Mitsutsuji & Yamakage, 2020) or the "influence layer" (Zino et al., 2020a; 2020b). The difference is that the two different dynamics happen sequentially (Mitsutsuji & Yamakage, 2020) or simultaneously (Zino et al., 2020a, b) in the previous models, but in SNOAEs, only one of the dynamics is allowed each time, depending on the network structure.

In the models of general behavior, the term "behavior" has an abstract meaning that is no longer limited to "expressing opinions". A common setting in these models is that opinions are continuous, usually in the range of 0 and 1, and behaviors are binary, usually taking the value of either 0 or 1 (e.g., Martins, 2008; Zino et al., 2020a, b; Zhan et al., 2021). The advantage of this setting is that opinions can be viewed as the preference for one of the two behaviors. With this simple behavior-opinion relation, observed behaviors are treated as "extreme opinions" that can be directly added to or subtracted from the opinions (e.g., Zino et al., 2020a, b; Zhan et al., 2021). As a result, these models do not require a "behavior-opinion inference process" as in the AOI model, making them fundamentally similar to the models of opinion expression. However, this advantage does not come without a cost: such a design limits the possibility of modeling other forms of behaviors whose relations with opinions might be more complex and indirect.

5.2.3. SUMMARY: BEHAVIOR AS A MESSENGER OF OPINION

The lesson we learned from the literature is that, surprisingly, the core of "behavior-opinion co-evolution" may not be related to "behavior" itself. As opinions are supposed to be private and unobservable, we need a kind of messenger that can bring our opinions (whether truthfully or not) to others. In turn, we also receive messengers from others and translate them back into opinions (whether accurately or not). So the real factor that is directly interacting with and hence exerting influence on our opinions is the translated or inferred opinions of others that are brought to us by the messenger. The messenger may come up with various forms – in classic opinion dynamics models, the messenger is private opinion itself; in models of opinion expression, it is expressed opinion/ the behavior of expressing opinions; in CODA, AOI, CMAO, and SNOAEs, it is an abstract behavior. Therefore, what is really coupled to opinion dynamics is the dynamics of messenger – how the messenger is created based on private opinion, how the messenger is received and translated by its receiver, and how the translated messenger – in the form of inferred opinion – affects the receiver's opinion. The reason that "behavior" or "action" is used in the name of existing models (e.g., CODA, AOI) and the framework that is going to be proposed in Section 5.3 is because it is probably the most common messenger of

opinions in daily life.

So, if most messengers are behaviors, why do we need this "messenger theory" after all? Indeed, the theory provides the guiding philosophy for building the framework: behavior is a messenger of opinion. This philosophy is helpful when making decisions about whether a type of interaction or mechanism should be included in the framework. For example, the direct influence of one's opinion on other people's behavior is impossible as there lacks an intermediate messenger, but a direct interaction between behavior to behavior is allowed as one's behavior requires no messenger to interact with others' behaviors. Second, the theory/ philosophy would constantly remind us of two important mechanisms: how opinion determines behavior (i.e., how the messenger is created by the sender), and how behavior is translated into opinions (i.e., how the messenger is read by the receiver). These two mechanisms are key to the co-evolution but are also easily forgotten without this theory in mind, especially when the relation between opinion and behavior is simple and direct (see CODA, CMAO, and SNOAEs).

5.3. THE FRAMEWORK

In this section, we present the framework MOBEP (**M**odeling **O**pinion-**B**ehavior Co-Evolution for **E**xplaining Opinion **P**olarization) in detail. We start with the axioms, followed by key components of the framework, and then discuss their relations, together with other components that are necessary for constructing a complete framework. To validate the framework, a selected number of models are decomposed accordingly. Finally, we summarize the functionalities and contributions of the framework.

5.3.1. AXIOMS

The MOBEP framework is built upon a small number of axioms derived from the literature that have already been mentioned explicitly or implicitly in the paper:

- An agent's opinion is not observable by other agents;
- An agent's behavior is observable by the agent's interacting partners; and
- An agent's behavior serves as the messenger of her opinion.

The last axiom has been explained in Section 5.2.3. Note that in classic opinion dynamics models as well as Mitsutsuji & Yamakage (2020), Zino et al. (2020a, b), and Zhan et al. (2021), in some cases, agents can directly exchange opinions via communication. However, as stated in Section 5.2.3, there is actually an implicit behavior in the communication – expressing one's opinion truthfully.

5.3.2. KEY COMPONENTS

According to the so-called "messenger theory" (Section 5.2.3), there are three types of possible interactions in the co-evolution of opinions and behaviors: (i) personal opinion-behavior interaction (one's opinion affects her own behavior), (ii) interpersonal behavior-behavior interaction (one's behavior affects another agent's behavior), and (iii) interpersonal behavior-opinion interaction (one's behavior affects another agent's opinion). The

following interactions are not possible because an opinion needs a corresponding behavior to interact with other factors: (i) interpersonal opinion-opinion interaction (one's opinion affects another agent's opinion); (ii) interpersonal opinion-behavior interaction (one's opinion affects another agent's behavior). Meanwhile, personal opinion-opinion and behavior-behavior interactions (one's opinion/ behavior affects her own opinion/ behavior) may exist, depending on whether the dynamics is memoryless. They will be encoded in the component of schedule.

The interpersonal opinion-opinion interaction, which should be excluded from the framework according to the theory, is the major interaction in classic opinion dynamics models and also exists in other opinion-behavior co-evolution models (e.g., Mitsutsuji & Yamakage, 2020; Zino et al., 2020a, b; Zhan et al., 2021). Opinions of different people cannot interact directly, so what a classic opinion dynamics model describes is in fact a combination of interpersonal opinion-behavior interaction and interpersonal behavior-opinion interaction, but the implicitly embedded behavior is "disclosing opinion truthfully" with the condition that others understand it correctly. If we do exclude this type of interaction, there will be two major drawbacks. First, to model opinion dynamics in a conversation, where it is plausible to assume that the behavior is "disclosing opinion truthfully" if we don't consider any type of dishonesty or obfuscation, one would need to break this process into two processes repeatedly. Second, the behavior of interest is usually not "disclosing opinion truthfully". Therefore, we will need to deal with two behaviors in one model, which is unnecessarily redundant. To avoid these drawbacks, we would like to include the interpersonal opinion-opinion interaction in the model, with a special note that it is a combination of two interactions with a special type of behavior.

The four types of interactions (personal opinion-behavior, interpersonal behavior-behavior, interpersonal behavior-opinion, interpersonal opinion-opinion) are the cornerstone of the opinion-behavior co-evolution, upon which the key components of the framework are built. The key components are listed as follows.

Opinion-driven behavior change answers "how one's behavior is affected by her opinion" (personal opinion-behavior interaction). In CODA, AOI, and SNOAEs, this is the only mechanism that determines an agent's behaviors. The simplistic choice, as in CMAO, is to model opinions as the preferences for certain behaviors. For example, let agent i's opinion $o_i \in [0,1]$ be the probability that she will choose behavior A instead of B. If agent i's behavior is solely determined by her opinion, she then chooses A with probability o_i, and chooses B with $1-o_i$. In the AOI model, behavior change is guided by behavior-opinion relations: an agent can only choose the behaviors that are obliged or permitted by her opinion.

If multiple behaviors are available according to the mechanism of opinion-driven behavior change, an intelligent agent would strategically choose one of them for certain purposes. For example, Tang et al. (2021a) consider a strategy called "obfuscation", meaning an agent would choose the behavior that discloses the least information about her opinion, probably motivated by the desire to protect privacy. This component also opens up the possibility of modeling similar concepts such as deception, ambiguity, spiral of silence (Noelle-Neumann, 1974), and pluralistic ignorance (Miller & McFarland, 1987).

Normative social influence answers "how one's behavior affects other's behavior" (interpersonal behavior-behavior interaction). In social psychology, "normative social

influence" and "informational social influence" are the two aspects of social influence (Deutsch & Gerard, 1955; Buechel et al., 2015). Normative social influence refers to "*an influence to conform with the positive expectations of another*", or simply "*conform behaviorally with the expectations of others*" (Deutsch & Gerard, 1955). The goal is to obtain "*social approval from others*" (Cialdini & Goldstein, 2004) instead of learning opinions, and therefore it only deals with behaviors, while opinions are not directly involved (if normative social influence is to change one's opinion instead of behavior, then the goal is not achievable because opinions are unobservable according to the axiom). To preserve the real world's heterogeneity, modelers also need to take into account outliers who refuse to conform behaviorally; that is, they are negatively affected or not affected by normative social influence (see the "counter-conformity" motive in Buechel et al. (2015) as an example).

In addition to the opinion dynamics literature, there is a considerable amount of theories and models describing relations between behaviors and opinions (alternatively called "beliefs", "intentions", "attitudes", etc[3]) in psychology and other fields, such as the theory of reasoned action (Fishbein & Ajzen, 1975), theory of planned behavior (Ajzen, 1991), and health belief model (Janz & Becker, 1984). For a particular topic, empirical studies can provide realistic insights into how opinion drives behavior (see Barile et al. (2021) for an example in the context of mask wearing). These types of works provide more options for modelers to implement the components of opinion-driven behavior change and normative influence.

Informational social influence, together with **behavior-opinion inference**, answers "how one's behavior affects another agent's opinion" (interpersonal behavior-opinion interaction). As mentioned above, informational social influence is another aspect of social influence, which refers to "an influence to accept information obtained from another as evidence about reality" (Deutsch & Gerard, 1955). In the context of opinion-behavior co-evolution, it means that an agent obtains information (opinion) from the behaviors of others to create or modify her own opinion (Packer et al., 2021). For example, when we see most people are wearing face masks during a pandemic, we may think that "face masks must be effective in preventing infection otherwise there would not be so many people wearing them". Consequently, we might develop a pro-mask opinion. The goal of informational social influence is to "form an accurate interpretation of reality and behave correctly" (Cialdini & Goldstein, 2004), which translates into, in our context, updating one's opinion according to the inference of others' opinions based on their behaviors. Here the goal of "truth-seeking" is modified to "opinion updating", as models of opinion dynamics usually do not care about "truth" or "reality". In the practice of modeling, informational social influence depends on the "behavior-opinion inference" process to translate observed behaviors into opinions. This process is the reverse of "opinion-driven behavior change": the opinion is first encoded into a behavior by its holder ("opinion-driven behavior change"), and the behavior is decoded back to an opinion once observed by others ("behavior-opinion inference"). Similar pairs of concepts have been identified in many disciplines such as cultural studies (e.g., Hall, 2007) and biology (e.g., Paninski et al., 2007; Purvis & Lahav, 2013). The reverse process usually suffers from a loss of

[3]These terms do have subtle differences in psychology, but we choose to ignore them, because "opinion" is used in a generic way here to represent an unobservable characteristic of agents.

accuracy because of two intrinsic reasons. First, information loss has already occurred in the "opinion-driven behavior change" process. For example, in the CODA model, it is impossible to obtain the exact opinion underlying a behavior as opinions are continuous but behaviors are discrete or even binary. Second, behaviors can also be affected by another agent's behavior through normative social influence. Observers may not take into consideration this extra influence, and therefore cannot make an accurate inference. In fact, the discrepancy between one's true opinion and the opinion inferred from her behavior is at the core of opinion-behavior co-evolution – if these two opinions are always the same, then the co-evolution reduces to classic opinion dynamics. In practice, the inference process can be modeled by a number of well-developed techniques such as Bayesian learning (e.g., Gale & Kariv, 2003; Acemoglu et al., 2011) and the beta-binomial model (e.g., Khalvati et al., 2016, 2019). In the AOI model, the inference process is to calculate the posterior distribution of opinions in the agent's neighborhood given each opinion's probability of leading to the observed behavior [4]. In some cases, the agent's own opinion also affects the inference process by serving as a reference (e.g., Banisch & Olbrich (2019). See Section 5.3.5 for details).

Direct opinion influence answers "how one's opinion affects another agent's opinion" (interpersonal opinion-opinion interaction), which is the fundamental question in classic opinion dynamics models. Despite all the variants, there are two major types of direct opinion influence in the literature, namely positive and negative influence. Positive influence assumes that agents would become similar (in terms of opinions) after interactions. For example, an agent would adopt the opinion of a random neighbor (e.g., the voter model, see Clifford & Sudbury (1973), Holley & Liggett (1975), and Krapivsky et al. (2010)), the opinion of the majority (e.g., the majority rule model, see Galam (2002)), or a weighted average of her and her neighbors' opinions (e.g., the DeGroot model, see DeGroot (1974)). Negative influence, on the contrary, proposes that interactions between sufficiently dissimilar people will make them even more dissimilar. In particular, negative influence is implemented via the so-called "influence weight" (Flache et al., 2017) that adjusts both the strength and valence of the influence: if the opinion difference between the interacting agents exceeds a certain threshold, the weight becomes negative and so does the influence (Macy et al., 2003; Flache & Macy, 2011; Mäs et al., 2014; Feliciani et al., 2017).

The mechanisms of direct opinion influence are also applicable to informational influence as the latter can be viewed as a direct opinion dynamics between one's private opinion and the inferred opinion. Conceptually they are two distinct components, but practically they share the same toolkit.

5.3.3. FRAMEWORK STRUCTURE

The framework structure is summarized in Figure 5.1. The key components are organized into two types of dynamics: direct opinion dynamics and behavior dynamics. In the first type where direct communication happens, an agent's opinion will be directly affected by another agents' opinion through **direct opinion influence**. This situation is called **direct opinion dynamics**, where the adjective "direct" stresses that no intermediate messenger (i.e., behavior) is involved. In the second type, agents can only observe

[4] The AOI model ignores prior probabilities by assuming no information is known before observation.

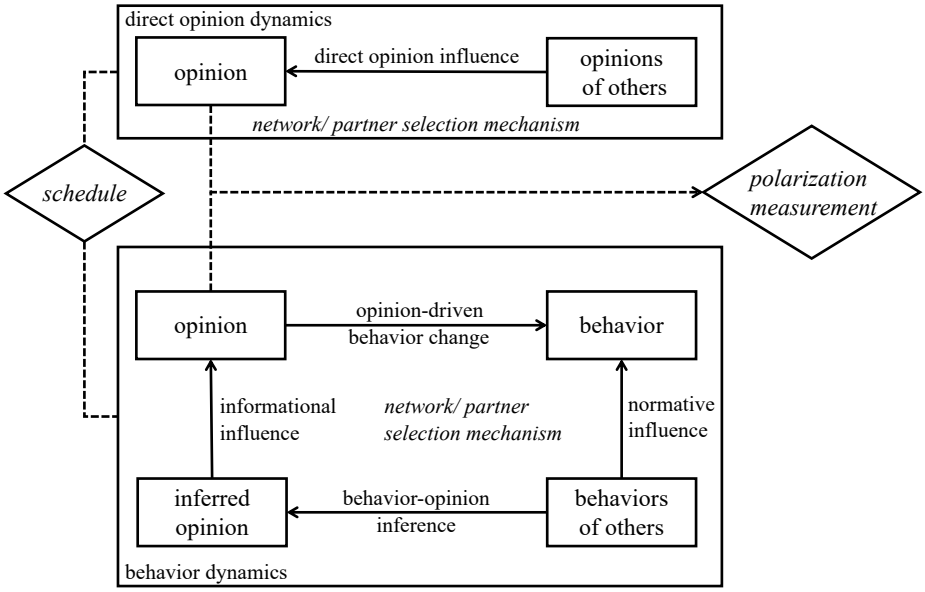


Figure 5.1: Structure of the MOBEP framework.

other agents' behaviors and cannot directly communicate with them. The following mechanisms will be working. An agent's behavior is jointly determined by her opinion through **opinion-driven behavior change** and the observed behavior of others through **normative influence**. Meanwhile, the observed behavior is translated into inferred opinion via **behavior-opinion inference**, and the inferred opinion affects the agent's opinion via **informational influence**. There is no rule telling us whether normative influence or informational influence happens first. It depends on whether the agent thinks before she behaves. This situation is called "behavior dynamics" to signify the role of behavior, though opinions may also change during the process. In the second type of dynamics where direct communication happens, an agent's opinion will be directly affected by others' opinions through **direct opinion influence**. This situation is called **direct opinion dynamics**, where the adjective "direct" stresses that no intermediate messenger (i.e., behavior) is involved. This is because, in many real-life circumstances, behavior dynamics and direct opinion dynamics jointly shape our opinions. For example, Mitsutsuji & Yamakage (2020) propose that people exchange honest opinions about war with intimate friends (i.e., direct opinion dynamics) and express (probably) different opinions in public (i.e., behavior dynamics).

It is quite obvious that the two types of dynamics are in nature the same as the dynamics in Mitsutsuji & Yamakage's private and public spheres (Mitsutsuji & Yamakage, 2020), or Zino et al.'s communication and influence layers (Zino et al., 2020a, b) respectively. In fact, the first type (direct opinion dynamics) represents classic opinion dynamics, and the second type (behavior dynamics) represents the opinion-behavior co-evolution

described by the CODA and AOI model. According to our messenger theory (Section 5.2.3), the first type can be safely included in the second type by defining the behavior as "expressing opinions truthfully", but we make the distinction deliberately in order to highlight the necessity of considering both types of dynamics in modeling opinion-behavior co-evolution.

Each dynamics requires a *social network* or *partner selection mechanism* to determine who interacts with whom in which order. In the tradition of opinion dynamics studies, agents that are directly connected by at least one edge/link of the social network are called neighbors, and usually, only neighbors are allowed to interact with each other (see an exception in Zhan et al. (2021)). Partner selection is a more dynamic process because agents can select their own neighbors according to certain rules such as homophily (Mäs & Flache, 2013; Mäs & Bischofberger, 2015). In the MOBEP framework, direct opinion dynamics and behavior dynamics should have different networks or partner selection mechanisms as they may occur between different types of partners. For example, concerning mask wearing, behavior dynamics usually occur between local strangers (e.g., those who shop at the same supermarket) but direct opinion dynamics are more likely to occur between friends, families, or global strangers (e.g., those who discuss mask wearing on the Internet).

The two dynamics are connected through an implementation component called *schedule*, which tells us "what happens in what order". In our context, it particularly refers to the order in which the two dynamics take place. For example, it may take two direct opinion dynamics for one (both numbers are arbitrary) behavior dynamics to happen. The schedule component determines the frequency and hence the relative strength of each dynamics's effect on agents' opinions. More generally, the component may also describe others issues such as which agent should be selected to update her opinion first, and whether sequential, parallel, asynchronous, or synchronous updating should be used (Wilensky & Rand, 2015).

Finally, the component *polarization measurement* measures the degree of opinion polarization of the system. The choice of polarization measurement reflects the modeler's understanding of polarization, and is essential to the results. One of the most used polarization measurements in the agent-based modeling and opinion dynamics community would be the FM index (Flache & Mäs, 2008; Flache & Macy, 2011). This component will be further discussed in Section 5.4.4.

5.3.4. APPLYING THE FRAMEWORK TO EXISTING MODELS

To validate the MOBEP framework, we apply it to selected models mentioned in Section 5.2. Regardless of the notations in the original papers, we use o for (private) opinions and m for behaviors or expressed opinions (except the CODA model).

Buechel et al.'s model: as a representative of models of opinion expression, Buechel et al.'s (2015) model of conformity can be decomposed as follows according to the framework:

- **Opinion:** a continuous variable o_i about a topic for agent i.
- **Behavior:** expressing one's opinion to the public. The stated opinion is m_i for agent i.

- **Opinions of others:** not observable.
- **Behaviors of others:** the stated group opinion q_i observed by agent i is the weighted average of her neighbors' opinions: $q_i = \sum_{j \neq i} \frac{g_{ij}}{1-g_{ii}} m_j$, g_{ij} is the weight between agent i and j.
- **Opinion-driven behavior change & normative influence:** the two components work jointly to determine one's behavior in an utilitarian manner. Agent i's behavior m_i is determined by maximizing $u_i = -(1-\delta_i)(m_i - o_i)^2 - \delta_i(m_i - q_i)^2$, where $\delta_i \in (-1, 1)$ is the conformity preference of agent i. The first term $-(1-\delta_i)(m_i - o_i)^2$ shows the role of opinion-driven behavior change, and the second terms $-\delta_i(m_i - q_i)^2$ shows the role of normative influence.
- **Behavior-opinion inference:** the stated opinion of agent j, m_j, is directly interpreted as the (inferred) opinion.
- **Inferred opinion:** the same as the stated opinion.
- **Informational influence:** the inferred opinion (i.e., the stated opinion) m_j affects agent i's opinion via the dynamics of the DeGroot model: $o_i = g_{ii} o_i + \sum_{j \neq i} g_{ij} m_j$.
- **Direct opinion influence:** does not exist.
- **Network/ partner selection mechanism:** for behavior dynamics, a weighted and directed network G is used.
- **Schedule:** as there is no direct opinion dynamics, schedules between direct opinion dynamics and behavior dynamics do not exist. Within behavior dynamics, at each time step, one or more agents are selected to state and update her/ their opinion(s).
- **Polarization Measurement:** the study is not particularly for polarization, so there is no polarization measurement.

CODA models: the CODA model (Martins, 2008) can be decomposed as follows according to the framework:

- **Opinion:** an agent i has a continuous probability $p_i \in [0, 1]$ that $m_i = 1$ is better than -1. In practice, its log odd form $v_i = ln(p_i/(1-p_i))$ is used.
- **Behavior:** an agent i needs to decide her behavior $m_i \in \{-1, 1\}$.
- **Opinions of others:** not observable.
- **Behaviors of others:** an agent i's neighbor j has the behavior $m_j \in \{-1, 1\}$ which is observable to agent i.
- **Opinion-driven behavior change:** $m_i = sign(p_i - 0.5)$ (Martins, 2014).
- **normative influence:** does not exist.

- **Behavior-opinion inference:** if $m_j = 1$, the inferred opinion will be $\Delta v_i = ln(\alpha/1-\alpha)$, and otherwise $\Delta v_i = -ln(\alpha/1-\alpha)$. $\alpha > 0.5$ is the probability (believed by agent i) that agent j will choose $m_j = 1$ ($m_j = -1$) if she prefers $m_j = 1$ ($m_j = -1$).
- **Inferred opinion:** Δv_i.
- **Informational influence:** v_i is updated to $v_i + \Delta v_i$.
- **Direct opinion influence:** does not exist.
- **Network/partner selection mechanism:** for behavior dynamics, the interaction rules of the voter model and the Sznajd model are used in the original paper (Martins, 2008).
- **Schedule:** there is no inter-dynamics schedule. Within behavior dynamics, the model follows the schedule of the voter model or the Sznajd model.
- **Polarization Measurement:** Martins (2008) shows a bimodal opinion distribution, and there is no specific polarization measurement.

AOI model: the AOI model (Tang & Chorus, 2019) can be decomposed as follows according to the framework:

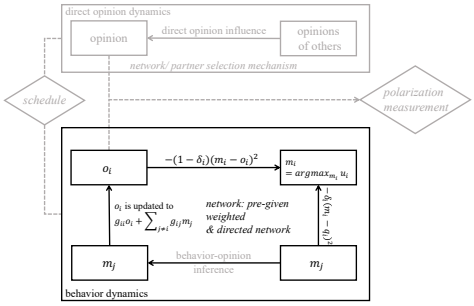
- **Opinion:** a discrete variable o_k. An agent i may take one of the possible opinions at a time: $o^{(i)} \in \{o_k\}_{k=1,...,K}$.
- **Behavior:** a discrete variable m_g (called "action"). An agent i may take one of the possible actions at a time: $m^{(i)} \in \{m_g\}_{g=1,...,G}$.
- **Opinions of others:** not observable.
- **Behaviors of others:** agent i's neighbor j has the action $m^{(j)} \in \{m_g\}_{g=1,...,G}$, which is observable to agent i.
- **Opinion-driven behavior change:** agent i determines her action according to the action-opinion relations. She can only take the action that is permitted or obliged by her opinion $o^{(i)}$. If her opinion permits more than one action, she chooses one of them randomly.
- **Normative influence:** does not exist.
- **Behavior-opinion inference & inferred opinion:** it is called "action-opinion inference". Agent i infers the opinion of agent j from her action $m^{(j)}$ in a Bayesian manner. Namely, agent i thinks agent j has the opinion o_k with the probability $P^{(i)}(o_k|m^{(j)}) = \frac{P(m^{(j)}|o_k)}{\sum_{z=1}^{K} P(m^{(j)}|o_z)}$ (which describes the inferred opinion).
- **Informational influence:** agent i will update her opinion according to the popularity of each opinion in her neighborhood; that is, she will take o_k with probability $\hat{P}^{(i)}(o_k) = \frac{\sum_{j \in M_i} P(o_k|m^{(j)})}{\sum_{z=1}^{K} \sum_{j \in M_i} P(o_z|m^{(j)})}$, where M_j is the collection of neighbors of agent i.

- **Direct opinion influence:** does not exist.
- **Network/ partner selection mechanism:** for behavior dynamics, the network is described by the Von Neumann neighborhood.
- **Schedule:** as there is no direct opinion dynamics, no schedule exists between direct opinion dynamics and behavior dynamics. Within behavior dynamics, at each time step, an agent is randomly selected to update her opinion and action.
- **Polarization measurement:** Tang & Chorus (2019) show the opinion distribution, and there is no specific polarization measurement.

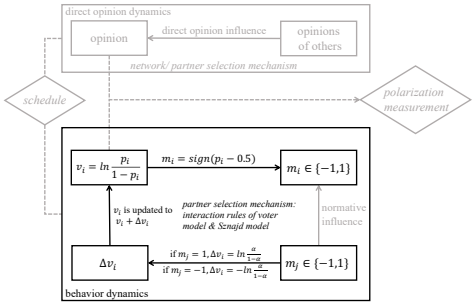
SNOAEs: Zhan et al. (2021)'s SNOAEs model can be decomposed as follows according to the framework:

- **Opinion:** agent i has the opinion $o_i \in [0,1]$.
- **Behavior:** agent i has the behavior $m_i \in \{0,1\}$.
- **Opinions of others:** whether the opinions of others are observed depends on the social network (see **Schedule**).
- **Behaviors of others:** whether the behaviors of others are observed depends on the social network (see **Schedule**).
- **Opinion-driven behavior change:** if $o_i < h_i$, $m_i = 0$; otherwise $m_i = 1$. h_i is agent i's feature (called "opinion threshold of action choice").
- **Normative influence:** does not exist.
- **Behavior-opinion inference:** does not exist.
- **Inferred opinion:** the same as m_j.
- **Informational influence:** the mechanism of bounded confidence model. If $|o_i - m_j| \leq \epsilon_i$ for some threshold (called "bounded confidence") ϵ_i, o_i is updated to $o_i + \alpha(m_j - o_i)$, $\alpha \in (0, 0.5]$; otherwise o_i remains unchanged.
- **Direct opinion dynamics:** the same as informational influence except that m_j is replaced by o_j as agents can directly exchange opinions.
- **Network/ partner selection mechanism:** both dynamics share the same social network.
- **Schedule:** at each time step, an agent chooses another agent as her interacting partner randomly. If the interacting partner is connected with the agent, they directly exchange opinions (i.e., direct opinion dynamics); otherwise the agent observes the behavior of her partner (i.e., behavior dynamics).
- **Polarization measurement:** the model is not designed for polarization studies, and there is no polarization measurement.

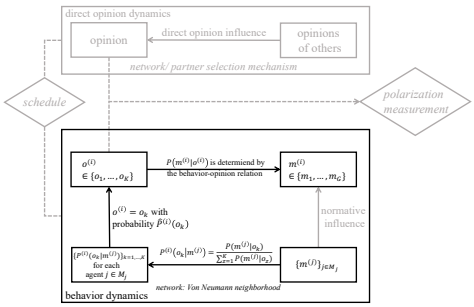
The fittings of the models to the framework are presented in Figure 5.2, which also shows one of the functionalities of the framework – organizing literature (see Section 5.3.5):



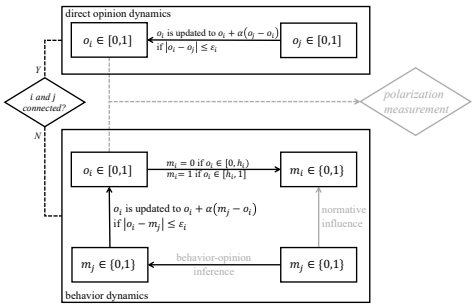
(a) Buechel et al.'s model of conformity



(b) CODA model



(c) AOI model



(d) SNOAEs model

Figure 5.2: Fitting selected models to the MOBEP framework. Framework components that do not appear in the model are marked gray. The schedule that is not relevant to the interactions of the two dynamics is not given in the Figure due to space limitations. Special note for (d): the social network is the same for both dynamics, which is a directed, unweighted social network (see Zhan et al. (2021) for details).

5.3.5. FUNCTIONALITIES OF THE FRAMEWORK

All the components of the MOBEP framework are borrowed or derived from existing theories and models, so how can this framework contribute to the burgeoning literature of opinion dynamics? Like most conceptual frameworks, the functionalities are related to both existing efforts and future studies.

Organizing existing models: After decades of development, the literature of opinion dynamics and polarization has grown into a huge collection of mathematical and computational models. Because of the varieties in terminology and barriers between sub-fields, it is usually difficult to acknowledge relations between models and find their places in the literature. For example, models of opinion expression (see Section 5.2.1) may not be placed in the same category as the CODA, AOI, or CMAO because they usually use terms such as "stated opinion" (Buechel et al., 2015), and "overt attitude" (Mitsutsuji & Yamakage, 2020), rather than "behavior" or "action". However, after decomposing them into framework components, we can find that models of opinion expression share basic ideas with CODA, AOI, and CMAO to a non-trivial extent. To further demonstrate this

functionality, we use the framework to test if the persuasion model (Mäs & Flache, 2013) and social feedback model (Banisch & Olbrich, 2019) are in fact describing similar dynamics as the opinion-behavior co-evolution in models of opinion expression or models of general behavior (e.g., CODA, AOI, CMAO, SNOAEs). The persuasion model, in the simplest way, assumes that opinion is formed by various arguments. Agents are more likely to interact with others with similar opinions (i.e., the mechanism of homophily), and will adopt the arguments of their interacting partners (Mäs & Flache, 2013). First, because opinion is formed by arguments, "opinion" in the model should be the "behavior" in the framework, and "argument" in the model should be the "opinion" in the framework. To avoid confusion, here we use the same terms as in the model. Although opinion takes part in the partner selection mechanism, it does not serve as the messenger of arguments because agents can directly learn the arguments of others. Therefore, the model does not fit into the behavior dynamics component. As a matter of fact, the persuasion model can be encompassed by the direct opinion dynamics, where an agent's arguments are affected by her partner's arguments, while the partner selection mechanism is related to the similarity between agents' opinions. As a result, although the persuasion model also features the duality of opinions, it is fundamentally different from models of opinion expression or general behavior as reviewed in Section 5.2, and is more similar to classic opinion dynamics models [5].

The social feedback model (Banisch & Olbrich, 2019), on first inspection, is not related to opinion-behavior co-evolution. It assumes that an agent expresses her opinion (which is assumed to be binary) to a randomly selected neighbor, and the neighbor will give her feedback. Basically, if the neighbor has the same opinion as the agent, the feedback is positive, and otherwise negative. By accumulating feedback, the agent develops an internal evaluation of each opinion, and in the next time step, she will express the opinion that has the better internal evaluation. If we consider "opinion" in the model as the "behavior" in the framework, and consider "internal evaluation" in the model as the "opinion" in the framework, the social feedback model can fit well into the behavior dynamics component (see Figure 5.3). From Figure 5.3 we can see that in essence, the social feedback model is similar to the AOI model besides one major difference – the agent's own "behavior" (i.e., "opinion" in the social feedback model) also takes part in the behavior-opinion inference process in the former. This exception also tells us that the framework (in fact, any framework) is not a strict rule that cannot be disobeyed, and any modification that suits the particular model should be accepted.

Facilitating future polarization research: The ultimate goal of the framework is to advance opinion polarization studies in the context of opinion-behavior co-evolution. On the one hand, the framework is expected to bring attention to the role of behavior in opinion polarization; on the other hand, the framework should be able to facilitate future behavior-related opinion polarization studies by providing guidelines. Not all components in the framework need to be included in a new model, but their existence would keep asking the modeler, "will this component make the model more realistic/reasonable?". For example, many models either focus on "direct opinion dynamics" (e.g., classic opinion dynamics models) or "behavior dynamics" (e.g., CODA, AOI, and model

[5] Certainly, the persuasion model itself is a novel contribution to the literature. We claim its similarity with classic models from the perspective of opinion-behavior co-evolution.

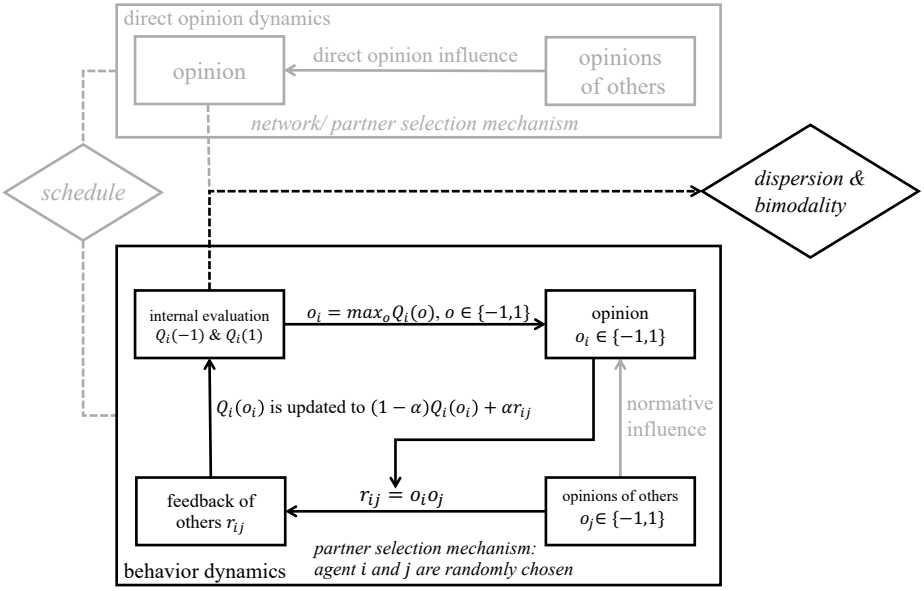


Figure 5.3: Fitting the social feedback model to the MOBEP framework. Framework components that do not appear in the model are marked gray. Additional information: $o \in \{-1,1\}$, and o_i can take one of the values of o. α is the learning rate.

of social feedback), but in some cases, it is of both theoretical and empirical importance to include both dynamics (e.g., Mitsutsuji & Yamakage, 2020). In addition, by comparing components and their relations, new models can be distinguished from existing models and hence avoid duplication and find their places in the literature.

5.3.6. COMPARING WITH OTHER FRAMEWORKS

The MOBEP framework can help us locate models in the literature of models, while we also need to locate the MOBEP framework in the literature of frameworks. There are not many frameworks of opinion dynamics, let alone frameworks of opinion-behavior co-evolution. We find the following frameworks that are worth mentioning and comparing with ours.

The unified framework for opinion dynamics (Coates et al., 2018) decomposes a typical opinion dynamics model into four "rule modules", including: (i) "structural rule" that describes the initial settings of the model such as population, opinion distribution, and social networks; (ii) "communication rule" that determines "who interacts with who" (Coates et al., 2018); (iii) "update rule" that determines how people's opinion is affected by other people's opinion; and (iv) "co-evolutionary rule" that determines if and how the social network changes during the dynamics of opinions. It is not difficult to find similar concepts in our MOBEP framework. For example, the communication rule is related to our partner selection mechanism, and the update rule is related to our direct

opinion influence. However, Coates et al.'s framework is developed on the basis of opinion dynamics models without behavior. Therefore, it does not highlight the role of behavior in the dynamics, and is not perfectly compatible with opinion-behavior co-evolution.

Together with CMAO, Zino et al. (2020a, b) also provide a relevant modeling framework. Basically, their framework assumes that one's opinion is affected by the opinions of others shared on the communication layer and the behaviors of others observed on the influence layer, while one's behavior is affected by her own opinion and the behaviors of others observed on the influence layer. Zino et al.'s framework is quite similar to the MOBEP framework, such as the "communication layer" in Zino et al.'s framework is essentially the same concept as "direct opinion dynamics" in the MOBEP framework. It is not surprising to find such similarities given that there are in total only four factors (one's own opinion, one's own behavior, other's opinion, and other's behavior) to be taken into consideration. The difference is that the MOBEP framework highlights the "behavior-opinion inference process", but this part is ignored in Zino et al.'s framework, which is best represented by its corresponding model, CMAO. In CMAO, the behaviors of others are directly treated as extreme opinions without any inference or translation. This is certainly reasonable if the opinion-behavior relation is sufficiently simple, but in a more general setting, the absence of the behavior-opinion inference process would be unacceptable, as the observed behaviors of others must go through the behavior-opinion inference process in order to be translated into inferred opinions, and only an inferred opinion can directly affect one's own opinion (see Chapter 3 and 4 for examples). In addition, the MOBEP framework also emphasizes the importance of a schedule component that arranges the interactions between direct opinion dynamics (communication layer) and behavior dynamics (influence layer). As a result of these factors, the MOBEP framework can better accommodate models that involve relatively complex opinion-behavior relations, such as the AOI model and SNOAEs.

5.4. CASE STUDY: MASK WEARING DURING THE COVID-19 PANDEMIC

To give a demonstration of how the MOBEP framework works, we propose a model of both opinion and behavior of mask wearing during the COVID-19 pandemic. According to the framework, the model consists of two types of dynamics: the direct opinion dynamics where people exchange opinions about mask wearing, and the behavior dynamics where people decide whether to wear a face mask in a local public space (such as a supermarket). The two dynamics are relatively independent (in the sense that the two dynamics don't need to happen at the same time) but closely intertwined (in the sense that they jointly shape the opinion landscape), together representing a typical daily life in the pandemic where other activities such as working in the office and traveling are less probable.

5.4.1. GENERAL SETTING

We investigate a set of agents of size N. The behavior of agent i ($i = 1, 2, ..., N$) is denoted by $m_i \in \{0, 1\}$, where $m_i = 1$ represents wearing a mask, and $m_i = 0$ represents not wearing a mask. Agent i's opinion $o_i \in [0, 1]$ represents her preference for wearing a mask, which can be a result of agent i's perception of the effectiveness of face masks, the severity

and susceptibility of the disease and many other factors internal to the agent (Barile et al., 2021). The value of $o_i(t)$ implies the extent to which agent i would like to wear a mask without other constraints. Formally, we assume o_i represents the agent i's "ideal probability of wearing a mask", the probability of wearing a mask when she can choose freely in the sense that (i) she has access to face masks, (ii) she is able to wear a mask, and (iii) there is no social influence, or, the agent does not care about any social influence.

The schedule of the model is as follows. Initially at time step $t = 0$, each agent $i = 1, 2, ..., N$ is assigned an opinion $o_i(t = 0)$, and she will choose to wear a mask ($m_i(t = 0) = 1$) with probability o_i, and choose not to wear a mask ($m_i(t = 0) = 0$) with probability $1 - o_i$. Since $t = 1$, each time step t is divided into two consecutive half-steps: t_a and t_b. The opinions assigned at $t = 0$ will be inherited by the agents in $t_a = 1_a$; that is, $o_i(0) = o_i(1_a), \forall i$. At t_a ($t \geq 1$), agents are randomly selected to go through the "behavior dynamics" process in their local neighborhoods. Once selected, agent i needs to make a decision between wearing a mask ($m_i(t_a) = 1$) or not ($m_i(t_a) = 0$) based on her opinion $o(t_a)$ and other agents' behaviors (i.e., opinion-driven behavior change and normative influence). Meanwhile, agent i will update her opinion from $o_i(t_a)$ to $o_i(t_b)$ according to her inference of other people's opinions underlying their behaviors (i.e., behavior-opinion inference and informational influence). This dynamics will be described in Section 5.4.2. In the second-half t_b, agents bring their opinions to the Internet, where all N agents are free to interact with each other and update their opinions from $o_i(t_b)$ to $o_i((t+1)_a)$ accordingly. This dynamics will be described in Section 5.4.3.

Each of the two dynamics has its own network/ partner selection mechanism. For behavior dynamics, agents can only observe and be observed by their neighbors in the given network. The network used here – the small-world network – will be further discussed in Section 5.4.6. For direct opinion dynamics, which is supposed to occur on the Internet, agents can interact with everyone else, including those who are not directly connected with them via the social network. A particular partner selection mechanism – homophily – is used to select interacting partners, which will be described in Section 5.4.3.

5.4.2. BEHAVIOR DYNAMICS

At the first half of each time step t_a, each agent observes the behaviors of her neighbors[6] and decides whether to wear a mask once and only once in a sequential order, which is randomized for each half time step. The behavior dynamics, as described in the framework, contains four components – opinion-driven behavior change, normative influence, behavior-opinion inference, and informational influence.

OPINION-DRIVEN BEHAVIOR CHANGE & NORMATIVE INFLUENCE

Agents need to determine whether to wear a mask or not in the behavior dynamics. To exclude factors that are irrelevant to our model, we assume that (i) there is no mask mandate, and (ii) all agents have access to face masks and all agents are able to wear

[6] An agent always observes the last behaviors of her neighbors; that is, an agent at t_a may observe her neighbor's behavior at $(t-1)_a$ (if $t_a = 1$, this would be 0) if the neighbor has not yet been chosen during t_a. This can be rationalized by assuming that either agents keep their masks on/off until the next time step, or they remember everyone's most recent behavior.

masks (that is, we exclude the situations that some people are unable to wear masks due to health conditions). Following Buechel et al. (2015) and Olcina et al. (2018), agent i determines her "real" (in contrast to "ideal") probability of wearing a mask ($P(m_i = 1)$) by maximizing the following utility function:

$$U_i = -(1-\delta_i)(P(m_i = 1) - o_i)^2 - \delta_i(P(m_i = 1) - \overline{m}_i)^2 \tag{5.1}$$

where $\delta_i \in [-1, 1]$ is the conformity parameter that "displays the relative importance of the preference for (counter-)conformity in relation to the preference for honesty" (Buechel et al., 2015). Agents with a positive δ_i (i.e., conformist) are positively affected by normative influence while agents with a negative δ (i.e., anti-conformists) receive normative influence in the reverse direction. The behaviors of others are summarized by $\overline{m}_i$, which is the average behavior of the agents observed by agent i (which is not to be confused with the average behavior of agent i):

$$\overline{m}_i = \frac{1}{n_i} \sum_{j \in L_i} m_j \tag{5.2}$$

5

where L_i, the local neighborhood, is the set of all agents observed by agent i (i.e., the neighbors of agent i) in her local neighborhood (excluding i itself). $n_i = |L_i|$ is the size of L_i.

The utility function takes an additive quadratic expression inherited from Buechel et al. (2015). According to Olcina et al. (2018), it is the "standard way economists have been modeling conformity behaviors". The first term on the right-hand side of equation (5.1) is the so-called "intrinsic part" (Buechel et al., 2015), penalizing the behaviors that are not in accordance with one's opinion. In other words, it reflects to what extent the agent has fulfilled her "intrinsic desire". In the context of the framework, this term represents opinion-driven behavior change. The second term is the so-called "social part" (Buechel et al., 2015) that reflects the agent's distance to the group norm (represented by the average observed behavior $\overline{m}_i$), representing normative influence in the framework.

The utility function represents a trade-off between "what the agent truly wants to do" and "what others (implicitly) want the agent to do". This idea has been found in many studies of conformity. Besides Buechel et al. (2015), Olcina et al. (2018) use a similar utility function to study assimilation behaviors of migrants and consider the function represents "a tension between personal preference and coordination with peers". Though not using a utilitarian approach, Ellinas et al. (2017) adopt this idea by modeling the "conflicting dynamics" where agents try to find a balance between "cognitive consistency" and "social conformity". Similarly, in Constant et al. (2019)'s active inference model, agents select actions by optimizing the expected free energy, consisting of a "pragmatic value" referring to the "potential of fulfilling preferred outcome" and an "epistemic value" that relates to uncertainty reduction.

Unlike Buechel et al. (2015) and Olcina et al. (2018) where the behavior of an agent can be directly obtained by maximizing the utility function, in our model, an agent can only determine her **probability** of wearing a mask from the utility function. This setup introduces some degree of uncertainty and randomness to behavior dynamics. The reason we deviate from the deterministic approach is twofold. First, social scientists have constantly

emphasized that uncertainty/ randomness is a fundamental element underlying human behaviors. Constant et al. (2019) claim that "there is an intrinsic motivation for everything we do that is completely independent of expected utility". Sociological models have shown that even a small and rare deviation from the deterministic assumption is able to generate significant differences in the outcome (Mäs & Flache, 2013; Macy & Tsvetkova, 2015), and the ignorance of noises is "not an innocent simplification" (Macy & Tsvetkova, 2015). Moreover, Mäs concludes that "randomness is a crucial building block of theoretical models", and a deterministic model for individual behavior might be "not only unrealistic but also potentially misleading" (Mäs, 2018). Second, it is difficult for agents to obtain meaningful or accurate inferences of the opinions underlying deterministic behaviors. For example, if m_i is determined by maximizing some utility function $u(m_i)$, then observing $m_i = 1$ only tells the observer that $u(1) \geq u(0)$, and solving this inequality will only give the range for o_i instead of a point estimate. Conversely, the probabilistic assumption opens up the possibility of applying behavior-opinion inference techniques (such as Bayesian learning, to be discussed in the "Behavior-opinion inference" part of Section 5.4.2), giving agents more intelligence in the task of inferring opinions.

Finally, we would like to explore the conceptual relations between $P(m_i = 1)$, o_i, and $\overline{m}_i$, the three major inputs of the utility function. While $P(m_i = 1)$ is the "real" probability of wearing a mask when both opinion-driven behavior change and normative influence are present, o_i, apart from its definition as an opinion, represents the "ideal" possibility of wearing a mask when normative influence is absent (see Section 5.4.1). Meanwhile, because $m_i \in \{0, 1\}$, $\overline{m}_i$ is in fact the share of mask wearers in the neighborhood. If we treat all agents in L_i as one "representative" agent, $\overline{m}_i$ is then regarded as this agent's real probability of wearing a mask. To summarize, the three variables are all probabilities of mask wearing but under different conditions, which explains why in the utility function they can be linearly combined without any transformation.

BEHAVIOR-OPINION INFERENCE

As the framework indicates, after observing her neighbors in L_i, agent i obtains an inferred opinion of them through behavior-opinion inference. Here we use the beta-binomial model (Murphy, 2012; Khalvati et al., 2016, 2019) to describe the inference process. The core assumption is that agent i would treat all members in L_i as one single "average agent" or "representative agent" deciding whether to wear a mask n_i times (Khalvati et al., 2019). In real life, people usually don't care who (often strangers) is wearing a mask but focus on how many people are wearing masks, and therefore it is reasonable for an observer (i.e., agent i) to view everyone else as one agent. A similar argument has been given by Khalvati et al. (2016) that "individuals cannot be tracked by others, and all group members can be seen together as one group" (Khalvati et al., 2016).

Each time this "average agent" – as a representative of everyone in L_i – will wear a mask with probability $P(\overline{m}_i = 1)$, which will be denoted as $\theta \in [0, 1]$ to save ourselves from complex notations. Therefore, in all n_i times, the probability that this average agent will wear a mask $n_i^m = \theta n_i$ (which is the number of agents in L_i wearing masks) times is:

$$P(n_i^m|\theta) = \binom{n_i}{n_i^m} \theta^{n_i^m} (1-\theta)^{n_i - n_i^m} \tag{5.3}$$

$P(n_i^m|\theta)$ can also be interpreted as the probability that n_i^m out of n_i average agents choose to wear masks in one time step.

Suppose agent i's prior belief of θ follows a Beta distribution:

$$P(\theta) = Beta(\alpha, \beta) = \frac{\theta^{(\alpha-1)}(1-\theta)^{(\beta-1)}}{\int_0^1 \theta^{(\alpha-1)}(1-\theta)^{(\beta-1)}\,d\theta} \tag{5.4}$$

where $\alpha, \beta > 0$ are the so-called "hyper-parameters" of the prior distribution (Murphy, 2012). Choosing the Beta distribution as the prior is a common practice as it is a conjugate prior of Binomial distribution (see Murphy (2012), and Khalvati et al. (2019)). According to Bayes' theorem, after observing the behaviors of her neighbors (i.e., the "average agent"), agent i will have the following inference of θ:

$$P(\theta|n_i^m) = \frac{P(n_i^m|\theta)P(\theta)}{P(n_i^m)} = \frac{P(n_i^m|\theta)P(\theta)}{\int_0^1 P(n_i^m|\theta)P(\theta)\,d\theta} \tag{5.5}$$

5

Through simple calculation, we know:

$$P(\theta|n_i^m) = Beta(\alpha + n_i^m, \beta + n_i - n_i^m) \tag{5.6}$$

Comparing equation (5.4) and equation (5.6), it is easy to find that "the posterior is obtained by adding the prior hyper-parameters to the empirical counts" (Murphy, 2012).

In the next time step when agent i visits the local public place to experience behavior dynamics again, she will use the posterior from the last time step as her current prior. Without any specific knowledge, at $t = 1$ (the first step when people observe others' behaviors)[7], we don't include any prior but set $P(\theta|n_i^m) = Beta(n_i^m, n_i - n_i^m)$ (Khalvati et al., 2019). As a result, at time step $t^* \geq 1$, her posterior will be:

$$P(\theta|n_m(t^*)) = Beta(\sum_{t=1}^{t^*} \lambda^{(t^*-t)} n_i^m(t), \sum_{t=1}^{t^*} \lambda^{(t^*-t)}(n_i(t) - n_i^m(t))) \tag{5.7}$$

where $\lambda \in [0,1]$ is the decay rate (Khalvati et al., 2019). Note that $n_i^m(t)$ is the number of mask-wearers observed by agent i at time t (more precisely, t_a) and is not the number of mask-wearers at the end of t. Meanwhile, because we are using a static network, n_i, which is the size of agent i's neighborhood, does not change with t.

Given the posterior distribution of θ, we choose its mean as the inferred opinion (i.e., the estimate of θ by agent i):

$$\hat{\theta}(t^*) = \frac{\sum_{t=1}^{t^*} \lambda^{(t^*-t)} n_i^m(t)}{\sum_{t=1}^{t^*} \lambda^{(t^*-t)} n_i(t)} \tag{5.8}$$

We assume that the agents will take $\hat{\theta}$ as the inferred opinion of her neighbors. Technically it is easy to go one step further by taking into account normative influence and

[7] All the dynamics described in Section 5.4.2 happen during the first half of each time step including $t = 0$. It will be more confusing if the subscript a is included, especially for the equations. Therefore, in Section 5.4.2 only, the subscript is ignored.

estimating the average opinion of all agents in L_i from $\hat{\theta}$. However, such an extra step is considered unnecessary due to people's bounded rationality. In particular, the well-known concept of "fundamental attribution error" (Ross, 1977) describes an omnipresent tendency for humans to attribute observed behaviors or outcomes (here: wearing a mask) to dispositional or personal factors (here: opinion), even if the cause may be situational or environmental (here: conformity) (Heider, 1958; Harvey et al., 1981; Kosmidis, 2021). In our context, fundamental attribution error suggests that people are likely to underestimate or even ignore the effect of normative influence on other agents' behaviors.

INFORMATIONAL INFLUENCE

By "behavior-opinion inference", agent i obtains the inferred opinion $\hat{\theta}$, which is expected to affect agent i's opinion through informational influence. In the context of mask wearing, seeing many others wearing masks, even if it does not directly provide any argument, will inform the observer that "there must be some (unknown) reasons to wear masks otherwise there wouldn't be so many people wearing them".

We use the rejection model (Flache & Macy, 2011) – a representative of negative influence models (see Section 5.3.2) – for informational influence. Modifying the original implementation (Flache & Macy, 2011) for our context leads to the following opinion dynamics rules. The weight of the influence from neighbors to agent i at time step t_a, $w_i(t_a) \in [-1,1]$, takes the following expression:

$$w_i(t_a) = 1 - 2|\hat{\theta}(t_a) - o_i(t_a)| \tag{5.9}$$

The weight measures the similarity between o_i and $\hat{\theta}$ before informational influence and determines whether o_i will become more ($w_i > 0$) or less similar ($w_i < 0$) to $\hat{\theta}$ after informational influence. The "raw state change" (Flache & Macy, 2011) is then given by:

$$\Delta o_i(t_a) = w_i(t_a)(\hat{\theta}(t_a) - o_i(t_a)) \tag{5.10}$$

Finally, agent i updates her current opinion $o_i(t_a)$ to $o_i(t_b)$, the opinion that is going to be exchanged in direct opinion dynamics at t_b:

$$o_i(t_b) = \begin{cases} o_i(t_a) + \Delta o_i(t_a)(1 - o_i(t_a)) & \text{if } o_i(t_a) > 0.5 \\ o_i(t_a) + \Delta o_i(t_a) o_i(t_a) & \text{if } o_i(t_a) \leq 0.5 \end{cases} \tag{5.11}$$

Equation (5.11) is especially designed to assure that $o_i(t_b)$ will not exceed its range $[0,1]$.

5.4.3. DIRECT OPINION INFLUENCE/ DYNAMICS

At the second half of each time step t_b, agents go online and directly exchange opinions about face masks, which is the direct opinion influence/ dynamics component of the framework. This is modeled by N sequential events. In each event $\tau = 1, 2, ..., N$, an agent is randomly selected to update her opinion. Note that:

$$o_i(\tau = 0) = o_i(t_b) \tag{5.12}$$

$$o_i(\tau = N) = o_i((t+1)_a) \tag{5.13}$$

for all $i = 1, 2, ..., N$ if event τ happens in t_b. Agents are chosen without replacement; that is, all agents will be chosen once at this half time step, because we don't expect an agent's opinion to be influenced by online information more than once during this short interval.

Suppose in each event, the chosen agent i will be influenced by a set of other agents G_i ("global neighborhood"). Unlike the local neighborhood L_i that only includes agents that are neighbors of agent i, G_i can include anyone in the system. For simplicity, we set the size of G_i to be 1 in this model.

Agents determine whom to interact with according to homophily, referring to "greater interaction between like-minded individuals" (Dandekar et al., 2013). Homophily is one of the most substantial features of online communication due to the personalization techniques embedded in online services (Mäs & Bischofberger, 2015). Its relation with polarization (and hence the Internet) has been at the center of debates in the field though no consensus has been reached yet (e.g., Dandekar et al., 2013; Mäs & Flache, 2013; Mäs & Bischofberger, 2015). A common practice to implement homophily in opinion dynamics models is to let the probability that agent j is chosen by agent i take the following form (Mäs & Flache, 2013; Mäs & Bischofberger, 2015):

$$p(i,j) \propto \frac{(sim_{i,j})^h}{\sum_{k \neq i}(sim_{i,k})^h} \tag{5.14}$$

where $sim_{i,j}$ is the similarity between the opinions of agent i and j:

$$sim_{i,j} = 1 - |o_i - o_j| \tag{5.15}$$

Homophily then enters the dynamics through $sim_{i,j}$: agents whose opinions are more similar to o_i are more likely to be chosen by agent i.

Once selected, an agent will update her opinion by communicating with her online partner. Following informational influence, we again choose the rejection model (Flache & Macy, 2011) to simulate direct opinion dynamics. If agent i is selected at event τ to update her opinion, she will first choose one partner, say agent j, according to equation (5.14), and then the following dynamics would happen in a way similar to its counterpart in informational influence:

$$w_{ij}(\tau) = 1 - 2|o_j(\tau) - o_i(\tau)| \tag{5.16}$$

$$\Delta o_i(\tau) = w_{ij}(\tau)(o_j(\tau) - o_i(\tau)) \tag{5.17}$$

$$o_i(\tau+1) = \begin{cases} o_i(\tau) + \Delta o_i(\tau)(1 - o_i(\tau)) & \text{if } o_i(\tau) > 0.5 \\ o_i(\tau) + \Delta o_i(\tau) o_i(\tau) & \text{if } o_i(\tau) \leq 0.5 \end{cases} \tag{5.18}$$

The opinion of agent j is not affected.

If agent i is not selected in event τ, nothing happens to her:

$$o_i(\tau+1) = o_i(\tau) \tag{5.19}$$

5.4.4. POLARIZATION MEASUREMENT

A suitable measurement of polarization is an essential prerequisite for obtaining the results. Though there is no ideal measurement that everyone agrees upon (see Bauer (2019) for a review), the FM index (Flache & Mäs, 2008; Flache & Macy, 2011) would be one of the most popular choices among agent-based modelers. Basically, it is "the variance of the pairwise opinion differences between all pairs of agents in the population" (Mäs & Bischofberger, 2015) that takes the following expression (Flache & Macy, 2011):

$$FM = \frac{1}{N^2}\sum_{i,j}(d_{ij} - \bar{d})^2 \tag{5.20}$$

where d_{ij} is the opinion difference between agent i and j: $d_{ij} = 2|o_i - o_j|$, and $\bar{d}$ is the mean of all d_{ij}, including the case of $i = j$[8]. The prefactor 2 in the expression of d_{ij} is added in order to ensure that d_{ij} is within $[0,2]$ so that FM will be in the range of $[0,1]$[9]. $FM = 1$ indicates maximal/ perfect polarization where half agents take the opinion of 0 and the rest take 1, while $FM = 0$ indicates a consensus has been achieved.

Recently Tang et al. (2021b) proposed a novel polarization measurement, called Equal Size Binary Grouping Measurement (ESBGM)[10], to tackle the long-standing problem that a group structure based on the difference in the variable of interest (here: opinion) – which is fundamental in conceptualizing polarization – is usually absent when measuring polarization. The idea of the measurement (in the context of opinion dynamics) is rather simple: the population will be divided into two non-overlapping groups of equal sizes according to the similarities of their opinions. ESBGM is then defined as a function that increases with the between-group heterogeneity and decreases with within-group heterogeneity. Since our model only deals with opinions of one dimension, applying ESBGM becomes much easier as we only need to divide the population by the median value of opinions. Denote the two groups as C_1 and C_2 respectively, and define the centroid of each group C_k ($k = 1,2$), μ_k, as the average opinion of all agents in C_k, then the between-group heterogeneity is:

$$B = (\mu_1 - \mu_2)^2 \tag{5.21}$$

and the within-group heterogeneity of C_k is:

$$W_k = \frac{1}{N}\sum_{i \in C_k}(o_i - \mu_k)^2 \tag{5.22}$$

The condition $i \in C_k$ means agent i is in the group C_k. ESBGM takes the following expression:

$$ESBGM = \frac{B}{W_1 + W_2 + 1} \tag{5.23}$$

It is then easy to verify that $ESBGM = 1$ at perfect polarization ($B = 1$, $W_1 = W_2 = 0$), and $ESBGM = 0$ at consensus ($B = 0$, $W_1 = W_2 = 0$).

[8] In Flache & Macy (2011)'s paper, it is suggested that self-distance (the case of $i = j$) can be either excluded or included. If excluded, the maximal FM would be $1 - 1/(N-1)^2$; otherwise it will be 1.

[9] In the original papers (Flache & Mäs, 2008; Flache & Macy, 2011) there is no prefactor before $|o_i - o_j|$ because in their model settings, opinion $o_i \in [-1,1]$.

[10] The measurement is not named in Tang et al. (2021b).

To investigate the relation between these two measurements, we test them in the following situations. Suppose there are 50 agents: X of them, denoted as G_1, have the opinion of Y, while the rest $50-X$ agents, denoted as G_2, have the opinion of $1-Y$. Figure 5.4 shows the value of FM and ESBGM under various combinations of X and Y. From these examples, we can see that FM is in general less sensitive to X (representing the balance of the group[11] size) compared to ESBGM, especially under large X and small Y. In other words, a system with large FM but small ESBGM is likely to have a large opinion difference (i.e., the opinions of agents are located in the two extremes of the opinion spectrum) and an uneven distribution of group sizes. It is doubtless that the similarity of group sizes contributes positively to polarization (Esteban & Ray, 1994; Gigliarano & Mosler, 2009), but FM and ESBGM put different weights on it, which somehow represents the two views of understanding/ measuring polarization, namely the extremeness tradition and the cluster tradition (Bauer, 2019).

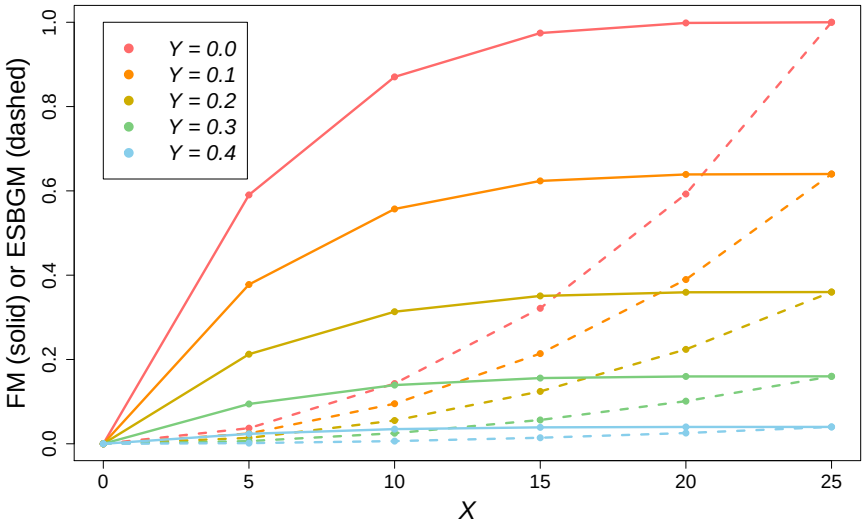


Figure 5.4: Comparison between FM (solid) and ESBGM (dashed).

In the rest of this section, we will present results in terms of both measurements to ensure that both views are taken into consideration. By paying the cost of being slightly redundant, we will obtain a more complete understanding of how relevant factors (e.g., conformity, homophily, etc.) affect opinion polarization.

5.4.5. FITTING MODEL TO THE MOBEP FRAMEWORK

Here we summarize the model in the context of the framework.

- **Opinion:** a continuous variable $o_i \in [0,1]$ about the preference (and the "ideal" probability) for wearing a mask by agent i.

[11] Here (and in the rest of this subsection) "group" should be defined as a collection of agents with relatively similar opinions. In the context of this example, it means G_k ($k = 1,2$).

- **Behavior:** a binary variable $m_i \in \{0, 1\}$ about whether to wear a mask ($m_i = 1$) or not ($m_i = 0$) for agent i.
- **Opinions of others:** an agent knows the opinion of her interacting partner in direct opinion dynamics.
- **Behaviors of others:** an agent (say agent i) can observe the behaviors of her neighbors in behavior dynamics, summarized by $\overline{m}_i$ (see equation (5.2)).
- **Opinion-driven behavior change & normative influence:** the probability that agent i wears a mask is determined by maximizing the utility function U_i (equation (5.1)), where the first term represents opinion-driven behavior change, and the second term represents normative influence.
- **Behavior-opinion inference & inferred opinion:** the beta-binomial model is used to translate observed behaviors into inferred opinions (see equation (5.8)).
- **Informational influence:** the inferred opinion affects agent i's opinion according to the rejection model (see equation (5.9)-(5.11)).
- **Direct opinion influence:** an agent directly communicates with her interacting partner, and her opinion is affected according to the rejection model (see equation (5.16)-(5.18)).
- **Network/ partner selection mechanism:** for behavior dynamics, agents can only interact with their network neighbors; for direct opinion dynamics, agents use the homophily mechanism to choose their interacting partners (see equation (5.14)-(5.15)).
- **Schedule:** one time step is divided into two half steps: in the first half step, all agents go through behavior dynamics sequentially and randomly; in the second half step, all agents go through direct opinion dynamics sequentially and randomly.
- **Polarization measurement:** both FM and ESBGM are used.

The fitting is also visualized in Figure 5.5. Unlike the models mentioned in Section 5.2, the case study contains all the framework components[12], thereby serving as a perfect demonstration of how the MOBEP framework works.

5.4.6. RESULTS

The model describes a relatively complex system where many interesting issues can be explored. In this subsection, we will focus on the conditions under which opinion polarization will emerge.

[12] Although CMAO also has most of the components, in its setup, the observed behaviors are directly interpreted as opinions, so we don't consider that it has a "complete" behavior-opinion inference process.

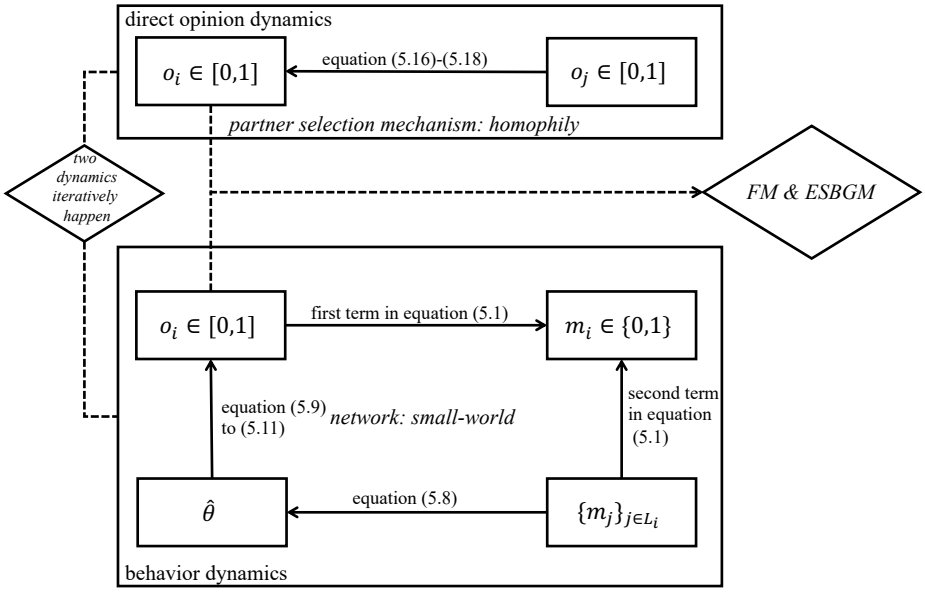


Figure 5.5: Fitting the case study model to the MOBEP framework.

BEHAVIOR DYNAMICS ALONE

We first run the behavior dynamics alone to see what will happen without direct opinion dynamics. We can frame this situation as a strict lockdown where people are only allowed to visit essential shops like supermarkets, and telecommunication, including the Internet, has not yet been invented or is not available. Among different combinations of conditions, opinion polarization would emerge on a small-world network composed of conformists.

A polarized case

We first present an extreme case when everyone has a maximum conformity preference (i.e., $\delta_i = 1, \forall i$). For graphical reasons as well as simplicity, we set the population size N to 50, and the total time step to 1000. We use the small-world network for behavior dynamics, which is a good representation of the daily situations during the pandemic (especially when a lockdown is imposed) where people's activities are restricted. The network is generated as follows: first, we create a so-called "disconnected caveman graph" describing a "maximally dense local network" (Flache & Macy, 2011). Each cave accommodates $s = 5$ agents, representing a small local community/ neighborhood or even a family (if a "stay-at-home order" is in place). In the disconnected graph, caves are isolated as there is no link between agents of different caves. Meanwhile, all agents in the same cave are connected. In the next step, each agent has a chance $p = 0.2$ to connect to another agent, avoiding duplication and self-connection. This step creates "shortcuts" between previously disconnected caves. The resulting network is not necessarily connected, as some caves may still be disconnected from others. However, as such disconnected caves

may lead to local consensus and hence significantly affect global polarization, we only choose the networks wherein all caves are connected. Such a network has a special property of "small-world", namely being highly clustered while enjoying relatively short distances between agents (Barrat et al., 2008). This property is in line with the fact that some people have connections outside their original "caves": they may visit several supermarkets, live on the boundary of two neighborhoods, or belong to two households (associated with a small average shortest path length), while others only visit the nearest supermarket, live in one neighborhood, or belong to one household in most of the time (associated with a high clustering coefficient). In this particular run, the small-world network is given in Figure 5.6, with an average clustering coefficient of 0.84 and an average shortest path length of 5.44898. Initially, each agent is randomly given an opinion $o_i \in [0,1)$[13]. Meanwhile, the decay parameter is set to be $\lambda = 0.5$.

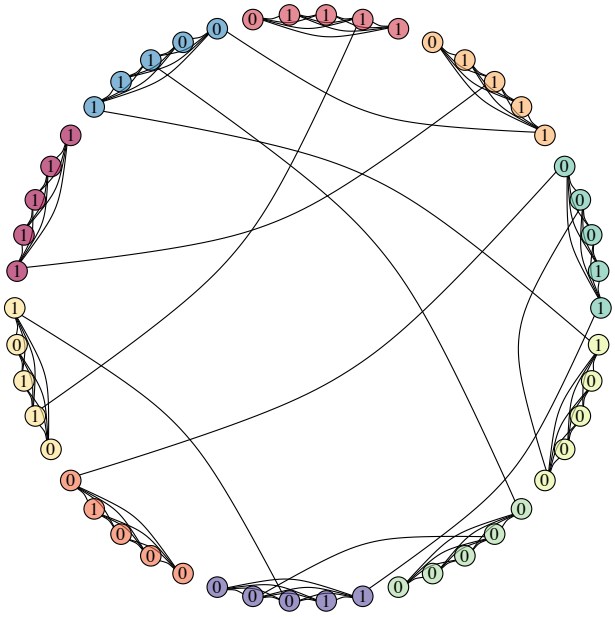


Figure 5.6: Opinion distribution on a small-world network at $t = 1000$ of a typical run with only behavior dynamics. A circle represents an agent, where the number inside is the agent's opinion, rounded to the second decimal place (the rounded results are either 0.00 or 1.00 so the parts after the decimal point are discarded to save space). Agents of the same color are in the same "initial cave" (the cave in the disconnected caveman graph) and are fully connected. Parameters include: $N = 50$, $s = 5$, $p = 0.2$, $\delta_i = 1$, $\forall i = 1, ..., N$, and $\lambda = 0.5$.

The final opinion polarization level is 0.9988112190139141 in terms of FM, and

[13]This is because in real coding, 1 is not included in most random functions. However, as the random function is still likely to take a very close value to 1, this exclusiveness can be ignored safely.

0.9992764309402817 in terms of ESBGM, both suggesting that the system is highly polarized in terms of opinions. The final opinion distribution on the network is also given in Figure 5.6, from which we can see that agents in the same cave do not always share similar opinions. This observation implies that the global polarization does not come from the opinion differences between caves. Although agents have polarized opinions, eventually all of them choose not to wear a mask[14]. The contradiction between "polarized opinions" and "unanimous behaviors" is, apparently, a result of high conformity.

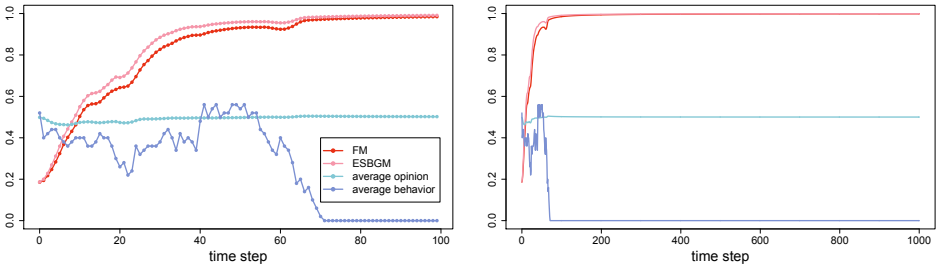
The dynamics of this particular run is described by the evolution of FM, ESBGM, average opinion, and average behavior (Figure 5.7). While average opinion remains relatively stable, opinion polarization levels (FM and ESBGM) rise dramatically in the first 100 time steps, and average behavior reaches zero (i.e., not wearing a mask) after some initial fluctuations.

The non-equilibrium feature of the dynamics

When the conformity parameter takes values other than one, the entire system may be out of equilibrium. By "out-of-equilibrium", we refer not only to the absence of an absorbing state where agents' opinions no longer change, but also to the irregular and unpredictable trajectories of the system's characteristics (i.e., polarization, average opinion, and average behavior) in a relatively long period. Figure 5.8 gives two examples of behavior dynamics that display the non-equilibrium feature. For such a complex system, finding an equilibrium state (if one exists) or a representative of "what the system finally ends up with" is at least an awkward (if not impossible) task that would probably lead us to statistical physics (see Krapivsky et al. (2010)). As a case study to illustrate the framework, the model does not aim to work out a "theoretical formalism" of the dynamics. Instead, we would like to study the general effect of relevant factors (such as conformity and homophily) on opinion polarization, which would probably be more meaningful in terms of explaining real-life phenomena and making policy suggestions.

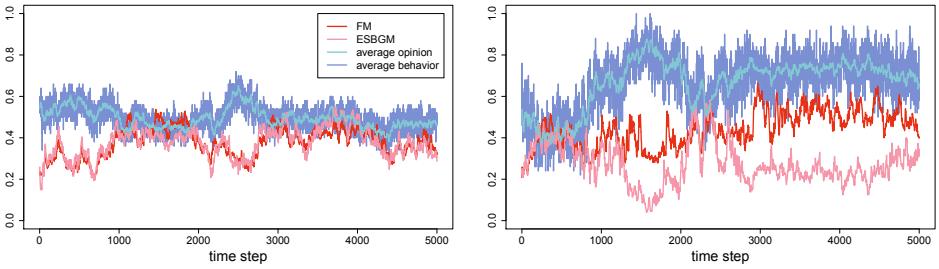
Very roughly, one time step in simulation is of a similar scale of one day in real life. Therefore, choosing an extremely large T (say, $T > 5000$) would be unrealistic considering that many relevant factors that are excluded from the model (such as the infection rate, development of vaccines, and mask wearing policies) may change dramatically in such a long period. Meanwhile, the dynamics under large conformity (say, $\delta_i = 1, \forall i$) won't be able to reach an equilibrium if T is too small (say, $T < 100$, see Figure 5.7). To achieve a balance between reality, efficiency, and computational cost, for the rest of this section, we would average each characteristic at $T = 1000, 2000, 3000, 4000$, and 5000 as a representative of the result.

[14] The result that no one wears a mask should not be generalized beyond this particular run: in other runs, it is likely to find out that everyone wears a mask.



(a) the first 100 steps (b) the whole process

Figure 5.7: Evolution of the FM, ESBGM, average opinion, and average behavior of a typical run (the run for Figure 5.6).



(a) $\delta = -0.6$ (b) $\delta = 0.6$

Figure 5.8: Typical trajectories of the behavior dynamics under moderate levels of conformity. Parameters include: $N = 50$, $s = 5$, $p = 0.2$, $\delta = -0.6$ (a) or 0.6 (b), and $\lambda = 0.5$. Note that they do not represent all the possible dynamics of the same parameters and conditions but are selected for illustration purposes only.

Conformity and opinion polarization in behavior dynamics

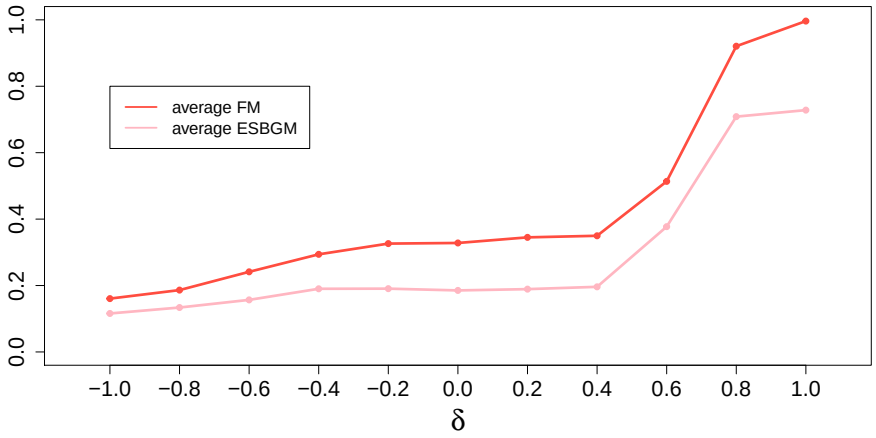


Figure 5.9: Polarization level (FM & ESBG) vs conformity preference (δ) in behavior dynamics. Average FM and ESBGM are obtained from the system at $t = 1000, 2000, 3000, 4000$, and 5000. Parameters include: $N = 50$, $s = 5$, $p = 0.2$, $\delta_i = \delta \ \forall i$, and $\lambda = 0.5$.

To see the effect of conformity on opinion polarization, we set $\delta_i = \delta$, $\forall i$ and run the simulation for different δ, ranging from -1 to 1. For each value of δ, we run 100 independent realizations while ensuring the network is connected. The result is graphically shown in Figure 5.9. It is clear that under given conditions, the polarization level (both FM and ESBGM) increases with everyone's conformity preference δ, especially when δ jumps from 0.4 to 0.6 and 0.6 to 0.8.

This conclusion is to some degree counter-intuitive as conformity is usually associated with consensus or uniformity instead of polarization. In fact, the high conformity preference does lead to a full/ approximate consensus in behavior. When $\delta_i = 1 \ \forall i$, the dynamics of behaviors is reduced to a voter model where agents take the behavior of a random neighbor (i.e., $P(m_i = 1) = \overline{m}_i$). The absorbing state of a voter model is always full consensus, even on small-world networks (Castellano et al., 2003; Castellano et al., 2009). Therefore, the behaviors of all agents will eventually be the same. In 100 trials[15] with the same setting (excluding initial opinions, behaviors, and networks) as in Figure 5.6 (also ensuring the network is connected), we obtain 51 replications of full consensus of wearing masks, and 49 replications of full consensus of not wearing masks, which is in line with what the voter model would predict.

To explain why opinions are polarized under behavioral conformity, one needs to take into account network topology. In the small-world network, each agent has a very small number (four if she does not know agents from other caves) of neighbors. Therefore, in

[15] which is the same experiment that generates part of Figure 5.9.

the initial part of the process, it is very likely that an agent's neighborhood is dominated by one behavior: in the case of four neighbors, this could be situations that all 4 neighbors or three out of four neighbors have the same behavior. Due to the clustering feature of the small-world network, the domination may survive for a sufficiently long time to drive each agent's opinion to one of the extremes. This is possible because the rejection model assumes that if agents have similar (different) opinions before interaction, the imposed social influence would be positive (negative). In our context, this means if an agent with an opinion against mask wearing (i.e., o_i is small) observes that the majority of her neighbors are wearing masks, she will develop an even more extreme anti-mask opinion (i.e., o_i decreases); on the contrary, if the agent wants to wear a mask and finds out that most or even all of her neighbors are wearing masks, her pro-mask opinion will become stronger. Given a sufficiently long period of domination, agents will develop very extreme opinions so that even if the domination of one behavior is overthrown (by the "external influence" from other caves via the shortcuts) and replaced by the other behavior, their opinions have already been anchored in extremes and can no longer be pulled back to moderate states. Because initially opinions are randomly initialized, the number of agents anchored in each extreme should be roughly equal, leading to a high level of global polarization. This theory is partly supported by Figure 5.7 as the average opinion is relatively stable (because agents' opinions are pushed towards one of the extremes in a roughly synchronous and symmetric manner) while FM and ESBGM increase rapidly.

To summarize, the simulation results of "opinion polarization and behavioral consensus" can be well explained by the two prominent features of the small-world network. The feature of "high local clustering" prevents the formation of opinion consensus and, together with other factors (rejection model, high conformity, etc.), generates opinion polarization, while the "small average distance" feature or the "shortcuts" fuels the formation of behavioral consensus.

COUPLED BEHAVIOR AND DIRECT OPINION DYNAMICS

Now we add the online communication (i.e., direct opinion dynamics in the framework) to the offline system (i.e., the behavior dynamics in the framework) to create an online-offline hybrid system. In particular, we are interested in the degree of polarization of the whole population under different levels of homophily. As discussed before, homophily is one of the major features of online communication. When there is no homophily ($h = 0$), each time an agent is influenced by another random agent. As homophily increases (i.e., h becomes larger), an agent is more likely to interact with like-minded others, and hence her opinion is more likely to be influenced by similar opinions of others.

According to Mäs & Bischofberger (2015), a rejection model with zero homophily ($h = 0$) would always generate polarized opinion distributions, and a larger level of homophily can significantly reduce polarization. Given that polarization increases with conformity in behavior dynamics (see Figure 5.9) and decreases with homophily in direct opinion dynamics, it would be interesting to investigate the joint effect of these two factors on opinion polarization in such a hybrid system.

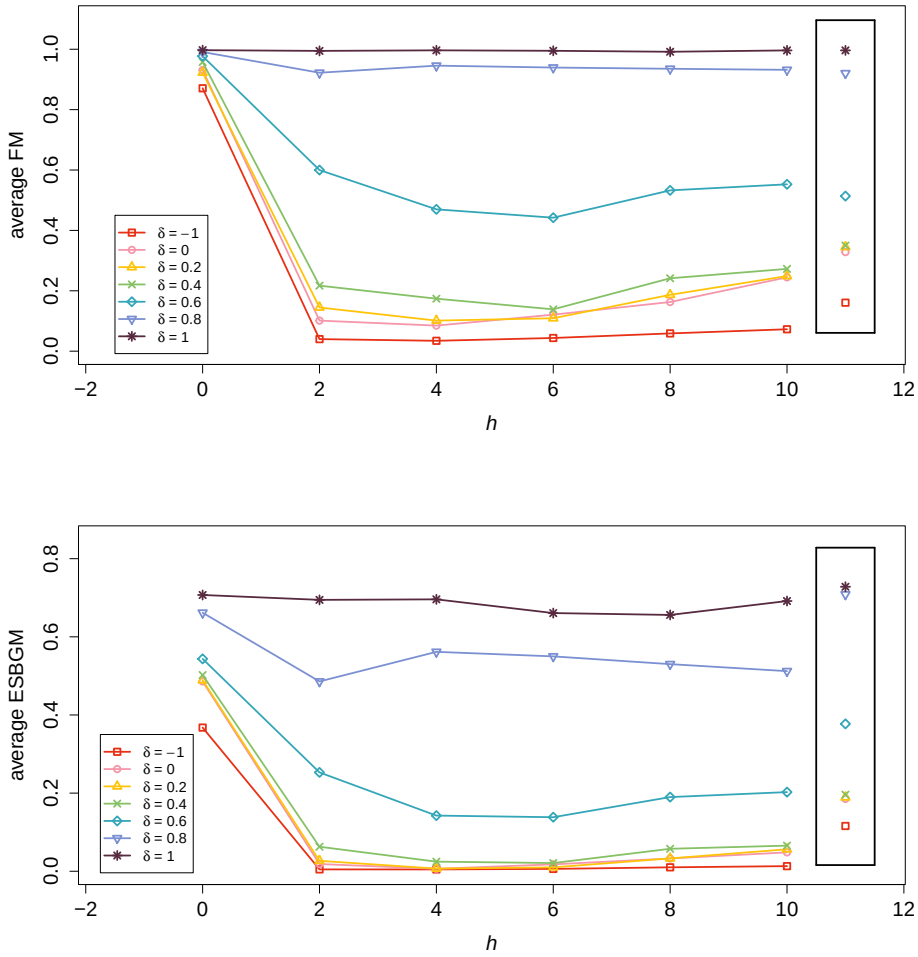


Figure 5.10: Effect of homophily h on opinion polarization measured by FM and ESBGM under different values of universal conformity preference δ averaged between $t = 1000, 2000, 3000, 4000$, and 5000. Points in the box on the right-hand side are the polarization levels in behavior dynamics only (i.e., without direct opinion dynamics) with corresponding δ, which are taken from Figure 5.9. Parameters include: $N = 50$, $s = 5$, $p = 0.2$, $\delta_i = \delta \; \forall i$, and $\lambda = 0.5$.

Figure 5.10 compares polarization levels of hybrid systems under various combinations of δ and h. In terms of FM, online communication (i.e., direct opinion dynamics) in the absence of homophily ($h = 0$) can significantly foster polarization of almost all offline systems (i.e., behavior dynamics) with different δ (except the offline system with $\delta = 1$ which has already reached almost full polarization without including online communication). The result implies that the driving force to polarization provided by direct opinion dynamics is strong enough to lead the hybrid system out of its non-equilibrium state.

As h increases to 2, all systems with $\delta \leq 0.6$ experience a sharp decrease in average FM. In particular, the polarization levels of systems with $\delta \leq 0.4$ are even smaller than their offline counterparts. These sharp decreases should be attributed to the fact that the direct dynamics alone (which is basically a rejection model with homophily) would generate low levels of polarization at $h = 2$: from $h = 0$ to $h = 2$, the driving force to polarization suddenly switches to a pushing force to consensus.

When we turn to ESBGM, the major observations in terms of FM are still valid, except one remarkable point: the polarization-fostering effect of zero homophily from direct opinion dynamics is less significant. For example, the average ESBGM of the system with $\delta = -1$ is only about 0.4 when $h = 0$, even if its average FM is more than 0.8. According to our previous analysis (Figure 5.4), the difference between FM and ESBGM suggests that the direct opinion dynamics without homophily is more powerful in creating extreme opinions than in creating groups of similar sizes.

It has become a tradition to blame the Internet for fostering opinion polarization. However, the case study shows that, at least in theory, the Internet may be wrongly accused. Conversely, behavior conformity may be an overlooked reason for polarization when taking into account behavior-opinion inference and informational influence. These conclusions may be premature, but undoubtedly they provide us with another possible approach to understanding opinion polarization, and remind us that opinion polarization is indeed a complex phenomenon.

5.4.7. ADDITIONAL NOTES

Although Section 5.4.6 has presented many interesting results, the case study is not intended to give a formal explanation of opinion polarization during the ongoing pandemic. Instead, its ambition is to illustrate how the framework can help us study opinion polarization. In addition, valuable lessons have been learned from the case study, and the most prominent one is that the co-evolution of opinion and behavior dynamics can be much more complex than direct opinion dynamics alone. In the co-evolution, behavior joins the game and brings not only more parameters but also higher degrees of stochasticity and uncertainty, which may lead to non-equilibrium dynamics that cannot be described deterministically (see Figure 5.8). However, like many other non-equilibrium dynamics, the co-evolution may still exhibit various statistical features when repeated many times.

Due to length limitations, many potentially interesting aspects of the case study remain unexplored. For example, agents are given the same conformity preference in our simulation to obtain a general idea of how conformity affects polarization. However, it would be both interesting and promising to investigate the situation where agents have different conformity preferences. In addition, there is a huge number of opinion dynamics models and social networks based on various theoretical assumptions, and testing the case study with different combinations of them is expected to provide us with a more complete understanding. However, digging into these aspects would fill another paper, so we leave these tasks for future studies.

5.5. SUMMARY AND DISCUSSION

In this paper, we present a unifying framework called MOBEP for studying opinion polarization in the coupled dynamics of opinions and behaviors. Guided by the intuition that "behavior is the messenger of opinion", the framework identifies five key components in the coupled dynamics, namely "opinion-driven behavior change", "normative influence", "behavior-opinion inference", "informational influence", and "direct opinion influence". The former four components form the so-called "behavior dynamics", and the last component forms "direct opinion dynamics". The two dynamics are connected by the component of schedule, and each of them has its own social network or partner selection mechanism. A number of models have been decomposed into these components in order to validate the framework. To further illustrate the framework, we have proposed and analyzed a case study of mask wearing during the ongoing COVID-19 pandemic, where all the framework components are well implemented. In particular, the case study shows that the coupled dynamics could be much more complex than direct opinion dynamics alone due to the introduction of behaviors. Therefore, re-examining existing mechanisms of polarization, which are usually derived from direct opinion dynamics, in the coupled dynamics shall provide new insights into our understanding of opinion polarization.

The MOBEP framework contributes to the opinion dynamics literature in different ways. First, it provides an architecture to organize existing opinion dynamics models. By decomposing them into relevant framework components, the hidden connections between models can be identified. Second, it facilitates future studies of opinion polarization in the coupled dynamics of opinions and behaviors by offering a general structure and relevant components. In particular, we would like to highlight a novel direction for future studies – using the framework to test the relations between different opinion dynamics mechanisms. Note that both "informational influence" and "direct opinion influence" are interpersonal opinion-opinion interactions, so if we implement them with competing mechanisms of (direct) opinion dynamics that will presumably lead to different polarization levels, the framework would then help us to test under what conditions one mechanism outperforms the other. Meanwhile, if the two mechanisms are known to generate similar polarization levels, with the framework we can test if the two mechanisms will reinforce or defer each other under different conditions (e.g., network, schedule, etc.).

Concerning the MOBEP framework itself, future works can improve its usability by building libraries for each framework component. Currently, we have introduced several candidate mechanisms for the components, such as obfuscation for "opinion-driven behavior change", and Bayesian learning for "behavior-opinion inference". Libraries of a sufficient number of such mechanisms would help modelers who are not familiar with the topic to implement each component easily by offering a wide range of choices. Once the libraries are constructed, the next step could be transferring this conceptual framework to practical tools such as a package for programming or a web-based application to simplify the modeling process.

REFERENCES

[1] Acemoglu, D., Dahleh, M. A., Lobel, I., & Ozdaglar, A. (2011). Bayesian learning in social networks. The Review of Economic Studies, 78(4), 1201-1236.

[2] Ajzen, I. (1991). The theory of planned behavior. Organizational Behavior and Human Decision Processes, 50(2), 179-211.

[3] Banisch, S., & Olbrich, E. (2019). Opinion polarization by learning from social feedback. The Journal of Mathematical Sociology, 43(2), 76-103.

[4] Barile, J. P., Guerin, R. J., Fisher, K. A., Tian, L. H., Okun, A. H., Vanden Esschert, K. L., Esschert, V., Jeffers, A., Gurbaxani, B. M., Thompson, W. W., & Prue, C. E. (2021). Theory-based behavioral predictors of self-reported use of face coverings in public settings during the COVID-19 pandemic in the United States. Annals of Behavioral Medicine, 55(1), 82-88.

[5] Barrat, A., Barthelemy, M., & Vespignani, A. (2008). Dynamical processes on complex networks. Cambridge University Press.

[6] Bauer, P. C. (2019). Conceptualizing and measuring polarization: A review. Working Paper. https://doi.org/10.31235/osf.io/e5vp8

[7] Buechel, B., Hellmann, T., & Klößner, S. (2015). Opinion dynamics and wisdom under conformity. Journal of Economic Dynamics and Control, 52, 240-257.

[8] Castellano, C., Fortunato, S., & Loreto, V. (2009). Statistical physics of social dynamics. Reviews of Modern Physics, 81(2), 591-646.

[9] Castellano, C., Vilone, D., & Vespignani, A. (2003). Incomplete ordering of the voter model on small-world networks. Europhysics Letters, 63(1), 153-158.

[10] Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. Annual Review of Psychology, 55, 591-621.

[11] Clifford, P., & Sudbury, A. (1973). A model for spatial conflict. Biometrika, 60(3), 581-588.

[12] Coates, A., Han, L., & Kleerekoper, A. (2018). A unified framework for opinion dynamics. In Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (pp. 1079–1086).

[13] Constant, A., Ramstead, M. J. D., Veissiere SPL, & Friston, K. (2019). Regimes of expectations: An active inference model of social conformity and human decision making. Frontiers in Psychology, 10, 679–679.

[14] Dandekar, P., Goel, A., & Lee, D. T. (2013). Biased assimilation, homophily, and the dynamics of polarization. Proceedings of the National Academy of Sciences, 110(15), 5791-5796.

[15] Deffuant, G., Neau, D., Amblard, F. & Weisbuch, G. (2000). Mixing beliefs among interacting agents. Advances in Complex Systems, 3(01n04), 87–98.

[16] DeGroot, M. H. (1974). Reaching a consensus. Journal of the American Statistical Association, 69(345), 118-121.

[17] Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. The Journal of Abnormal and Social Psychology, 51(3), 629-636.

[18] Ellinas, C., Allan, N., & Johansson, A. (2017). Dynamics of organizational culture: Individual beliefs vs. social conformity. PloS ONE, 12(6), e0180193.

[19] Esteban, J. M., & Ray, D. (1994). On the measurement of polarization. Econometrica: Journal of the Econometric Society, 62(4), 819-851.

[20] Feliciani, T., Flache, A., & Tolsma, J. (2017). How, when and where can spatial segregation induce opinion polarization? Two competing models. Journal of Artificial Societies and Social Simulation, 20(2).

[21] Fishbein, M., & Ajzen, I. (1975). Beliefs, attitude, intention, and behavior: An introduction to theory and research. Addison-Wesley.

[22] Flache, A., & Macy, M. W. (2011). Small worlds and cultural polarization. The Journal of Mathematical Sociology, 35(1-3), 146-176.

[23] Flache, A., & Mäs, M. (2008). How to get the timing right. A computational model of the effects of the timing of contacts on team cohesion in demographically diverse teams. Computational and Mathematical Organization Theory, 14(1), 23-51.

[24] 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).

[25] Gaisbauer, F., Olbrich, E., & Banisch, S. (2020). Dynamics of opinion expression. Physical Review E, 102(4), 042303.

[26] Galam, S. (2002). Minority opinion spreading in random geometry. The European Physical Journal B-Condensed Matter and Complex Systems, 25(4), 403-406.

[27] Gale, D., & Kariv, S. (2003). Bayesian learning in social networks. Games and Economic Behavior, 45(2), 329-346.

[28] Gastner, M. T., Oborny, B., & Gulyás, M. (2018). Consensus time in a voter model with concealed and publicly expressed opinions. Journal of Statistical Mechanics: Theory and Experiment, 2018(6), 063401.

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

[30] Gigliarano, C., & Mosler, K. (2009). Constructing indices of multivariate polarization. The Journal of Economic Inequality, 7(4), 435-460.

[31] 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 Multiagent Systems (pp. 1196–1204).

[32] Hall, S. (2007). Encoding and decoding in the television discourse. In A. Gray, J. Campbell, M. Erickson, S. Hanson & H. Wood (Eds.), CCCS selected working papers (vol. 2, pp. 402-414). Routledge.

[33] Harvey, J. H., Town, J. P., & Yarkin, K. L. (1981). How fundamental is "the fundamental attribution error"?. Journal of Personality and Social Psychology, 40(2), 346-349.

[34] Hegselmann, R., & Krause, U. (2002). Opinion dynamics and bounded confidence models, analysis, and simulation. Journal of Artificial Societies and Social Simulation, 5(3).

[35] Heider, F. (1958). The psychology of interpersonal relations. John Wiley & Sons.

[36] Holley, R. A., & Liggett, T. M. (1975). Ergodic theorems for weakly interacting infinite systems and the voter model. The Annals of Probability, 3(4), 643-663.

[37] Huang, C. Y., & Wen, T. H. (2014). A novel private attitude and public opinion dynamics model for simulating pluralistic ignorance and minority influence. Journal of Artificial Societies and Social Simulation, 17(3).

[38] Janz, N. K., & Becker, M. H. (1984). The health belief model: A decade later. Health Education Quarterly, 11(1), 1-47.

[39] Khalvati, K., Mirbagheri, S., Park, S. A., Drehere, J. C., & Rao, R. P. (2019). A Bayesian theory of conformity in collective decision making. In Proceedings of the 33rd International Conference on Neural Information Processing Systems (pp.9702–9711).

[40] Khalvati, K., Park, S. A., Dreher, J. C., & Rao, R. P. (2016). A Probabilistic model of social decision making based on reward maximization. In Proceedings of the 30th International Conference on Neural Information Processing Systems (pp. 2901-2909).

[41] Kosmidis, M. (2021). Fundamental attribution error. In M. Raz & P. Pouryahya (Eds.), Decision making in emergency medicine (pp. 153-158). Springer.

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

[43] Macy, M. W., Kitts, J. A., Flache, A., & Benard, S. (2003). Polarization in dynamic networks: A Hopfield model of emergent structure. In R. Breiger, K. Carley & P. Pattison (Eds.), Dynamic social network modeling and analysis: Workshop summary and papers. (pp. 162– 173). The National Academies Press.

[44] Macy, M., & Tsvetkova, M. (2015). The signal importance of noise. Sociological Methods & Research, 44(2), 306-328.

[45] Martins, A. C. (2008). Continuous opinions and discrete actions in opinion dynamics problems. International Journal of Modern Physics C, 19(04), 617-624.

[46] Martins, A. C. (2014). Discrete opinion models as a limit case of the CODA model. Physica A: Statistical Mechanics and its Applications, 395, 352-357.

[47] Mäs, M. (2018). The complexity perspective on the sociological micro-macro-problem. SSRN. https://ssrn.com/abstract=3129362

[48] Mäs, M., & Bischofberger, L. (2015). Will the personalization of online social networks foster opinion polarization? SSRN. http://ssrn.com/abstract=2553436

[49] Mäs, M., & Flache, A. (2013). Differentiation without distancing. Explaining bi-polarization of opinions without negative influence. PloS ONE, 8(11), e74516.

[50] Mäs, M., Flache, A. & Kitts, J. A. (2014). Cultural integration and differentiation in groups and organizations. In V. Dignum & F. Dignum (Eds.), Perspectives on culture and agent-based simulations (pp. 71-542). Springer.

[51] Mäs, M., Flache, A., Takács, K., & Jehn, K. A. (2013). In the short term we divide, in the long term we unite: Demographic crisscrossing and the effects of faultlines on subgroup polarization. Organization Science, 24(3), 716-736.

[52] Miller, D. T., & McFarland, C. (1987). Pluralistic ignorance: When similarity is interpreted as dissimilarity. Journal of Personality and Social Psychology, 53(2), 298-305.

[53] Mitsutsuji, K., & Yamakage, S. (2020). The dual attitudinal dynamics of public opinion: An agent-based reformulation of L. F. Richardson's war-moods model. Quality & Quantity, 54(2), 439-461.

[54] Murphy, K. P. (2012). Machine learning: A probabilistic perspective. The MIT Press.

[55] Noelle-Neumann, E. (1974). The spiral of silence a theory of public opinion. Journal of Communication, 24(2), 43–51.

[56] Olcina, G., Panebianco, F., & Zenou, Y. (2018). Conformism, social norms and the dynamics of assimilation. Institute of Labor Economics (IZA) Discussion Papers, No. 11436.

[57] Packer, D. J., Ungson, N. D., & Marsh, J. K. (2021). Conformity and reactions to deviance in the time of COVID-19. Group Processes & Intergroup Relations, 24(2), 311-317.

[58] Paninski, L., Pillow, J., & Lewi, J. (2007). Statistical models for neural encoding, decoding, and optimal stimulus design. Progress in Brain Research, 165, 493-507.

[59] Purvis, J. E., & Lahav, G. (2013). Encoding and decoding cellular information through signaling dynamics. Cell, 152(5), 945-956.

[60] 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.

[61] Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), Advances in experimental social psychology (pp. 173–220). Academic Press.

[62] Shang, Y. (2019). Resilient consensus for expressed and private opinions. IEEE Transactions on Cybernetics. https://ieeexplore.ieee.org/document/8850322

[63] Shepherd, P., & Goldsmith, J. (2020). A reinforcement learning approach to strategic belief revelation with social influence. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 10, pp. 13734-13735).

[64] 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.

[65] Stauffer, D. (2003). How to convince others? Monte Carlo simulations of the Sznajd model. In AIP Conference Proceedings (Vol. 690, No. 1, pp. 147-155). American Institute of Physics.

[66] Sznajd-Weron, K. (2005). Sznajd Model and Its Applications. Acta Physica Polonica B, 36(8), 2537-2547.

[67] Sznajd-Weron, K., & Sznajd, J. (2000). Opinion evolution in closed community. International Journal of Modern Physics C, 11(06), 1157-1165.

[68] 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).

[69] Tang, T., Ghorbani, A., & Chorus, C. G. (2021a). Hiding opinions by minimizing disclosed information: An obfuscation-based opinion dynamics model. The Journal of Mathematical Sociology. https://doi.org/10.1080/0022250X.2021.1929968

[70] Tang, T., Ghorbani, A., Squazzoni, F., & Chorus, C. G. (2021b): Together alone: A group-based polarization measurement. Quality & Quantity. https://doi.org/10.1007/s11135-021-01271-y

[71] Wilensky, U., & Rand, W. (2015). An introduction to agent-based modeling: Modeling natural, social, and engineered complex systems with NetLogo. The MIT Press.

[72] Ye, M., Qin, Y., Govaert, A., Anderson, B. D., & Cao, M. (2019). An influence network model to study discrepancies in expressed and private opinions. Automatica, 107, 371-381.

[73] Zhan, M., Kou, G., Dong, Y., Chiclana, F., & Herrera-Viedma, E. (2021). Bounded confidence evolution of opinions and actions in social networks. IEEE Transactions on Cybernetics. https://ieeexplore.ieee.org/document/9325877

[74] Zino, L., Ye, M., & Cao, M. (2020a). A two-layer model for coevolving opinion dynamics and collective decision-making in complex social systems. Chaos: An Interdisciplinary Journal of Nonlinear Science, 30(8), 083107.

[75] Zino, L., Ye, M., & Cao, M. (2020b). A coevolutionary model for actions and opinions in social networks. In 2020 59th IEEE Conference on Decision and Control (CDC) (pp. 1110-1115).