On the Methods for Explaining Polarization of Private and Unobservable Opinions
An opinion-behavior co-evolutionary approach

Tang, T.

DOI
10.4233/uuid:b4b02f3b-b622-4226-8f8c-be26e2c52428

Publication date
2022

Document Version
Final published version

Citation (APA)
Tang, T. (2022). *On the Methods for Explaining Polarization of Private and Unobservable Opinions: An opinion-behavior co-evolutionary approach*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:b4b02f3b-b622-4226-8f8c-be26e2c52428

ON THE METHODS FOR EXPLAINING POLARIZATION OF PRIVATE AND UNOBSERVABLE OPINIONS

AN OPINION-BEHAVIOR CO-EVOLUTIONARY APPROACH

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

ON THE METHODS FOR EXPLAINING POLARIZATION OF PRIVATE AND UNOBSERVABLE OPINIONS

AN OPINION-BEHAVIOR CO-EVOLUTIONARY APPROACH

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus Prof. dr. ir. T. H. J. J. van der Hagen,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op maandag 4 juli 2022 om 12:30 uur

door

Tanzhe TANG

Master of Science in Non-Equilibrium Systems: Theoretical Modelling, Simulation and Data-Driven Analysis
King's College London, Verenigd Koninkrijk
geboren te Suzhou, China.

Dit proefschrift is goedgekeurd door de:

promotor: Prof. dr. ir. C.G. Chorus
copromotor: Dr. A. Ghorbani

Samenstelling promotiecommissie bestaat uit:

Rector Magnificus	voorzitter
Prof. dr. ir. C.G. Chorus	Technische Universiteit Delft, promotor
Dr. A. Ghorbani	Technische Universiteit Delft, copromotor
Onafhankelijke leden:	
Prof. dr. T. C. Comes	Technische Universiteit Delft
Prof. dr. M. E. Warnier	Technische Universiteit Delft
Prof. dr. A. Flache	Rijksuniversiteit Groningen
Prof. dr. F. Squazzoni	Università degli Studi di Milano, Italië
Dr. P. W. G. Bots	Technische Universiteit Delft, reservelid

ISBN 978-94-6384-352-2

Printed in the Netherlands by Ridderprint

PREFACE

When I was working at TU Delft, one of my hobbies after hours of reading and writing was to pick up a book of Ph.D. thesis from the bookcase in the corridor and read its preface. Over the years, I have witnessed many Ph.D. candidates sharing their tears of joy and words of acknowledgment on paper. Naturally, I asked myself, "what if I am going to write the preface?". Now, when I really have the chance, I am asking a different question, "why shall I even write the optional preface given that I am already exhausted after finalizing the rest of the thesis?" After all, a preface is never the center of interest, unless it is used to study the candidate's personal social network (which, by the way, is a good research topic). However, when I checked the thesis again before sending it to the printing company, I felt awfully sorry if these people who helped me so much across the years were not well appreciated.

My heartfelt appreciation, undoubtedly, first goes to my promotor Caspar and daily supervisor Amineh. Caspar is my source of inspiration. His endless creativity and broad knowledge made my Ph.D. journey exciting and rewarding. Amineh led me to the broader domain of agent-based modeling, where I started to find my real interest. They have made me the researcher that I was dreaming of being. Apart from their guidance and support, I have benefited from everything – large or small – they did for me. I would especially acknowledge their mental support during my countless burnouts, without which my life would be less delightful.

Most of my working hours – at least before the pandemic – were spent in the office with my lovely teammates: Aemiro, Andreia, Nicolas, Teodora, and Tom. We were such a close group that I wish a reunion would be around the corner. Outside the office, everyone in TLO and ESS should be acknowledged (forgive me for not listing everyone's name here): it might be our chit-chat next to the coffee machine or your question raised in our colloquium that gives me a new thought.

During my Ph.D. days, I have also been helped by many scholars outside TU Delft. I am particularly grateful to Flaminio, Federico, Andreas, and Michael. I am also indebted to all the editors and referees who helped me improve the published papers (and later, the chapters of the thesis). Needless to say, I gratefully thank my doctoral committee members for reviewing and improving the thesis.

Finally, I wish to thank my family and friends for everything. You may never hear of my research topic, but you are my driving force to explore the unknown fields.

Tanzhe Tang
Bedum, June 2022

CONTENTS

1

INTRODUCTION

1

1.1. BACKGROUND

On 21 September 2021, Willem-Alexander, the King of The Netherlands, delivered his yearly speech to the Senate and the House where he concluded that "*This concern and uncertainty* (referring to the Dutch people's worry about the country and the world) *is exacerbated by the increasingly polarising tone of public debate, at both national and international level.*"[1] In fact, the concern and uncertainty of our current society is not only fueled by "a polarizing tone" but mostly an outcome of "polarization" itself – especially opinion polarization, which can be roughly interpreted as "the presence of groups which are internally homogeneous, externally heterogeneous, and of similar size" (Gigliarano & Mosler, 2009). The 2020 US election has presented a vivid demonstration of political opinion polarization between Democrats and Republicans, accompanied by its dreadful consequence. In 2021, the polarization between pro- and anti-vaccine groups has become the newest force that is tearing our post-pandemic society apart.

Beyond the King's speech, opinion polarization has become one of the most popular and prominent topics among scholars and the general public for good reasons – they are omnipresent and mostly harmful. From BREXIT in the UK to the rise of the far-right in the European continent, from debates on climate change to the "culture war" of Hawaiian pizza[2], opinion polarization can be found everywhere in human societies, and with the help of social media, it is highly likely that we will witness more polarized cases in the future.

Apart from its omnipresence, people's concern about opinion polarization mainly comes from its close relation with conflicts. Generally speaking, polarization itself (not necessarily opinion polarization but can also be the polarization of income, religion, or ethnicity) is considered as a "particularly relevant correlate of potential or open social conflicts" (Esteban & Ray, 1994), and when it comes to opinion polarization, its consequence could be less predictive (and hence more pernicious) because opinions change much faster than other factors. As concluded by Harteveld (2021), "a little bit of polarisation is good, but too much endangers democracy". Opinion polarization indicates the decline of moderate opinions and the rise of extreme opinions, and the hostility between people of different political opinions is likely to cause various levels of conflicts such as the series of chaos before, during, and after the 2020 US election. Furthermore, many problems in our complex society require cooperation based on consensus, where opinion polarization is a major obstacle. At the moment of writing this thesis, I have started to see the gradual recovery of our world from the COVID-19 pandemic, which is a result of international cooperation between governments, scientists, pharmaceutical companies, and ordinary people with one consensus that science is the ultimate cure to the pandemic. Consider if half of the population support vaccines and the rest are anti-vaxxers, we may never be able to see the light at the end of the tunnel at this speed. As I am not an expert in biology or medicine, it is not my role to endorse or criticize any pro- or against-vaccine opinion related to COVID-19. However, if the society is polarized, then neither of the opinions can

[1] This is the official English translation of the speech. The original words in Dutch are "*Deze onrust en onzekerheid worden nog gevoed doordat het maatschappelijk debat nationaal en internationaal steeds vaker op polariserende toon wordt gevoerd.*" Both original and translated texts of the speech are available at https://www.koninklijkhuis.nl.

[2] https://www.economist.com/1843/2021/05/10/the-great-hawaiian-pizza-culture-war

be translated into practice easily or without a huge cost (such as the Dutch curfew riots in 2021).

Since the existence and consequence of opinion polarization have been identified, the next task is to find out why or how it happens. The mechanism of opinion polarization is of both academic interest and real value to the polarized and polarizing society. On the one hand, the existence of opinion polarization itself is perplexing to sociologists. A classic assumption in opinion dynamics is that people's opinions will become similar after interactions. Therefore, polarization is not an expected outcome (Mäs & Flache, 2013). As Abelson (1964) questioned "*what on earth one must assume in order to generate the bimodal outcome of community cleavage studies*", followed by Axelrod (1997) who asked "*if people tend to become more alike in their beliefs, attitudes, and behaviors when they interact, why do not all such differences eventually disappear*?" Therefore, explaining opinion polarization by a suitable mechanism has become an urgent theoretical task in sociology. On the other hand, the prerequisite for mitigating opinion polarization in real life is to first understand why or how it happens. For example, a popular conjecture is that social media has promoted opinion polarization because it creates filter bubbles where people only consume opinions that are similar to their own (Pariser, 2011; Spohr, 2017; Chitra & Musco, 2020). If this mechanism is valid, social media companies then should improve their algorithms (such as the algorithm for recommendation systems, see Dandekar et al. (2013)) so that their users will be exposed to more diverse opinions and polarization can hence be mitigated.

The importance of the task has motivated the birth of many possible mechanisms that can explain opinion polarization based on their particular assumptions and implementations in the past decades, despite the fact that most early opinion dynamics models usually lead to consensus (see Flache et al. (2017) for a review). These mechanisms can be roughly divided into three types. The first type of mechanism relies on the assumption of homophily, which refers to the tendency that people with similar opinions are more likely to interact with each other (McPherson et al., 2001; Mäs & Bischofberger, 2015). Given that interactions will make opinions similar, models of homophily, such as bounded confidence models (Deffuant et al., 2000; Hegselmann & Krause, 2002), can avoid consensus and generate moderate levels of polarization. The second type of mechanism assumes negative influences between dissimilar agents; that is, an agent's opinion will be reinforced after interacting with another agent whose opinion is sufficiently different from hers (Macy et al., 2003; Flache & Macy, 2011; Mäs et al., 2014). Compared to the first type, negative influence models can produce stable and higher (even maximum) levels of polarization but suffer from the lack of empirical support (Mäs & Flache, 2013; Takács et al., 2016). The last type of mechanism introduces additional factors which are usually related to the formation, expression, or exchange of opinions. For example, the persuasion model (Mäs & Flache, 2013) introduces "argument" as the "ingredient" of opinion in the sense that one's opinion is the average of all the arguments she has in mind. At each time step, a randomly selected agent (i.e., the focal agent) chooses an interacting partner according to how similar their opinions are (i.e., homophilous selection) and adopts a random argument from the partner. The homophilous selection process implies that the selected partner is very likely to have a similar opinion as the focal agent, and the newly-learned argument is therefore expected to support and reinforce the focal agent's

opinion, eventually leading to a polarized opinion distribution. Another example is the social feedback model (Banisch & Olbrich, 2019) that introduces agents' "internal evaluation" of two discrete opinions to control opinion expression. An agent gives positive feedback to an expressed opinion of another agent that is the same as hers, and negative feedback otherwise. According to the received feedback, an agent generates her internal evaluation of each opinion, and expresses the opinion with the better evaluation next time. Such a feedback-evaluation mechanism gives an "affective experience-based route to polarization"(Banisch & Olbrich, 2019).

Up to this point, the whole story seems rather complete: opinion polarization is important because of its omnipresence and harmful results, which motivates scholars to search for explanations, and luckily three types of opinion polarization mechanisms have been found. However, if we take a closer look, the mechanisms have mainly centered on "polarization" instead of "opinion": it seems that if we substitute "opinion polarization" with "behavior polarization" or simply "polarization", most mechanisms are still valid. In fact, the term "opinion" in the field of opinion dynamics and social influence (the two fields are largely overlapped with a blurred boundary) has a very generic meaning: as Axelrod (1997) defined, it is "what social influence influences", and according to Flache et al. (2017), the term "opinion" can also represent "belief", "behavior", and "attitude". Although it is difficult to find out why "opinion" was chosen at the very beginning, it is quite clear that existing mechanisms of "opinion" polarization, in most cases, are actually explaining polarization of "what social influence influences". To be specific, they explicitly or implicitly assume that opinions are observable and can be directly influenced by other people's opinions just like a "spin" or a "particle" in statistical physics (Krapivsky et al., 2010). The problem is, besides being "what social influence influences", opinions in real life are of fundamental difference from "spins" and "particles" in the sense that they are by nature private and unobservable unless told truthfully and inferred correctly.

The discrepancy between the meanings of "opinion" could cause complications. First of all, as discussed before, the omnipresence and pernicious consequence of opinion polarization are not only related to "polarization" but also equally related to "opinion": there is no need to worry about the polarization that "half people are men and half are women" (gender polarization) or "half people wear parfum and half wear Cologne" (fashion polarization). In short, the public, including policy makers, needs a mechanism of "opinion" polarization but academia is providing explanations of the polarization of "what social influence influences". Therefore, it remains uncertain if these mechanisms can indeed explain the particular polarization of "private and unobservable opinions". For policy makers, the uncertainty would become an obvious obstacle to translating these mechanisms into implementable policies as no one knows which part of "what social influence influences" will be affected besides opinions, and whether the opinion or other parts of "what social influence influences" will be affected. As a possible result, the process of mitigating opinion polarization could be delayed due to the lack of a "specific" understanding of "opinion" polarization. Second, the "spin nature" of opinion could limit opinion polarization studies (and more broadly, opinion dynamics studies) to a relatively abstract level, hampering the field's transition from statistical physics to social science. Statistical physics is one of the origins of opinion dynamics studies (Castellano et al., 2009; Krapivsky et al., 2010), and this explains why opinions in most models and mechanisms in

the field are fundamentally similar to spins and particles. Although ideas and tools from statistical physics have been widely used, opinion dynamics and opinion polarization are ultimately, or at least partly, within the social science domain as the goal is to understand our society better. If opinions are modeled in a similar fashion as spins or particles, the field will be more of a sub-division of many-particle systems than a discipline focusing on "opinion" in the sense of everyday usage[3].

To wrap up, this very first section has highlighted the following points:

- Opinion polarization has received ample attention from both scholars and the general public due to its omnipresence and pernicious consequence;
- A number of mechanisms have been proposed to explain opinion polarization, which can be roughly divided into three types, including homophily, negative influence, and introduction of additional factors;
- In most existing mechanisms, the term "opinion" has a very generic meaning which can be summarized as "what social influence influences". This is very different from our daily usage where opinions are "private and unobservable". As a result, in these mechanisms, opinions are observable and can be directly affected by other opinions;
- The discrepancy in the meanings of opinion could cause complications for both policy makers and scholars;

1.2. RESEARCH GOALS

To solve the complications described above and obtain a more specific and realistic understanding of opinion polarization, we need to narrow down the scope of "opinions" from "what social influence influences" to "private and unobservable opinions"[4], which is equivalent to transferring Abelson (1964) or Axelrod (1997)'s query to the following question:

How can the polarization of "private and unobservable opinions" be explained?

This is an open question with various potential answers depending on theoretical assumptions, technical implementations, and particular contexts. For example, the polarization in abortion attitudes may be explained quite differently from the polarization in the taste of Hawaiian pizza. Therefore, it is never my intention to give an ultimate answer or a mechanism that can explain the polarization of "private and unobservable opinions" of all kinds. Instead, my primary research goal of the thesis is:

- Primary Research Goal: To develop methods that could systematically support the exploration of mechanisms that can explain the polarization of "private and unobservable opinions".

[3] A detailed discussion about the field of opinion dynamics/ polarization can be found in Chapter 6.

[4] The reader may be curious that if opinions are private and unobservable, how can we empirically know opinion polarization exists? Practically, people use observable behaviors, such as voting and self-expressing, as reliable indicators of one's opinions. However, the interactions between opinions and behaviors are worth further exploration. See the discussion of sub-goal $2'$. I thank the doctoral committee for raising this question.

On first inspection, the research goal does not sound exciting and may disappoint those who simply want a mechanism to solve the complications. Despite the fact that an ultimate answer or mechanism may never exist, it is not particularly difficult to come up with one or two mechanisms that can partly answer the question from certain perspectives (which I implicitly did in Chapter 3). However, the experience from classic opinion dynamics studies (referring to the studies where opinions are "what social influence influences") implies that the problem will never be the lack of mechanisms but rather the overload of mechanisms (in the form of models) that "fail to identify how they add to insights of earlier work" (Flache et al., 2017). As a result, there will be a growing accumulation of mechanisms instead of an accumulation of insights (Flache et al., 2017). The methods are expected to mitigate this problem by not only facilitating the process of mechanism development but also providing guidance for further studies in order to coordinate individual efforts. I contend that compared to providing my own mechanisms that will be barely used by others, it is more valuable to make methodological contributions to this relatively new direction of opinion polarization studies.

The first prerequisite for any opinion polarization mechanism is a reliable polarization measurement. Polarization itself is a slippery concept in the sense that everyone has a somewhat similar impression about it but a universally accepted definition or measurement has never come to light. The lack of a universally accepted measurement further isolates individual studies by impeding the comparison between opinion polarization mechanisms. Another slippery aspect is that although most studies consider the notion of "group" (compared to "individual") – referring to the division of the population (data) based on their similarities – as a crucial ingredient in understanding polarization, a well-defined group structure has generally been missing in polarization measurements. The gap between how we understand and measure polarization corrodes our confidence in the reliability of existing measurements, and would consequently hamper the development of opinion polarization studies regardless of the meaning of the term "opinion". Considering the fundamental role of measurements in explaining opinion polarization, the first sub-goal of the thesis is:

- Sub-goal 1: To develop a formal and broadly applicable polarization measurement that is coherent with the notion of group.

Such a measurement is expected to be more reliable than existing ones as the widely accepted "group-based" understanding of polarization will then be fully captured. At the same time, its versatility (such as being able to deal with both uni- and multi-dimensional opinions) would promote cross-study comparison of opinion polarization mechanisms.

Compared to existing works of opinion polarization, the primary research goal of the thesis has particularly emphasized the "private and unobservable" feature of opinions. Therefore, the second sub-goal is:

- Sub-goal 2: To develop a modeling method that can incorporate the "private and unobservable" feature of opinions in opinion polarization mechanisms.

Everyday life tells us that because it is private and unobservable, an opinion needs to be expressed via a certain observable "messenger", from which the observers (receivers) of the messenger need to infer the opinion. The messenger can take various forms –

it can be as simple as a word, a speech, or, more frequently in daily life, a particular action. Considering that words and speeches also need to be "spoken" or "written", it is safe to give the messenger a more common name: "behavior". In turn, observing behaviors of others would also trigger changes in opinions. To summarize, the expression, transmission, inference, and change of opinions are all closely related to behaviors. In fact, the absence of behaviors in opinion polarization mechanisms and the ignorance of the "private and unobservable" feature of opinions are the two sides of one coin. Therefore, to transfer from the polarization of "what social influence influences" to the polarization of "private and unobservable opinions" (and hence solve the complications), it is essential to introduce "behavior" as the messenger of "opinion", which bears the great potential to arrive at a more realistic and specific mechanism of opinion polarization. Consequently, sub-goal 2 can then be translated into:

> Sub-goal 2′: To develop a modeling method that can incorporate behaviors in opinion polarization mechanisms as a messenger of opinions.

As sub-goal 2 and 2′ are fundamentally equivalent, I will collectively refer to them as sub-goal 2 hereafter.

As discussed before, the classic, behavior-excluded mechanisms are actually assuming that opinions are always expressed truthfully and inferred correctly (hereafter referred to as "default strategy"). Acknowledging the unique feature of opinions by introducing behaviors opens up the possibility of investigating topics that relate to the deviation from the "default strategy", such as deception, obfuscation, strategic disclosure, and misunderstanding, whose roles in opinion polarization are potentially crucial but are generally incompatible with behavior-excluded mechanisms. Therefore, a follow-up goal of sub-goal 2′, which also serves as the third sub-goal, is:

- Sub-goal 3: To develop a modeling method to study the effect of the deviation from the "default strategy", such as obfuscation, on opinion polarization.

Finally, it should be noted that explaining opinion polarization is a complex task that involves not only opinion dynamics but also statistical physics, sociology, network science, and agent-based modeling. Therefore, it requires interdisciplinary knowledge, techniques, and insights from previous works. In particular, there already exist a number of works that have incorporated behaviors in the context of opinion dynamics (e.g., Martins, 2008; Huang & Wen, 2014; Buechel et al., 2015; Sohn & Geidner, 2016; Mitsutsuji & Yamakage, 2020; Zino et al., 2020a, b; Zhan et al., 2021). Although their aims are mostly unrelated to opinion polarization, they have provided helpful insights about how to deal with the co-evolution of opinions and behaviors that future works of behavior-included opinion polarization mechanisms can benefit from. However, these works are scattered over various disciplines with different terminologies, and hence their insights have been rarely identified by the opinion polarization community. Considering all of these factors, the last sub-goal of the thesis is:

- Sub-goal 4: To develop a unifying framework of the co-evolution of opinion and behavior to organize existing efforts and facilitate future works of opinion polarization mechanisms.

The framework I intend to propose is expected to provide a general architecture of the co-evolution of opinion and behavior, based on which scholars can start to develop opinion polarization mechanisms in a more organized manner while exploiting the insights gained from existing works. More importantly, the framework is expected to help scholars identify the relations between their mechanisms and existing ones in order to locate them in the vast literature. This function could substantially increase the efficiency of insights accumulation by discouraging duplication of mechanisms and promoting acknowledgment of each other. In fact, when reviewing the development of opinion dynamics, many scholars have realized that it "has been uncoordinated and based on individual attempts" (Castellano et al., 2009), and hence appreciate the urgent need for a "general shared framework" (Castellano et al., 2009). I hope the framework can help avoid the detours made previously in opinion dynamics studies, and will be appreciated by future scholars when they review the development of the field.

The framework will also encompass the outcomes of the previous sub-goals as its components, namely the polarization measurement and two modeling methods, and hence concisely display the contributions of the entire thesis.

To summarize, the ultimate goal that arises from the current complications of opinion polarization studies is to acquire a more specific and realistic understanding of opinion polarization, where the term "opinion" is no longer broadly defined as "what social influence influences" but is defined by its "private and unobservable" feature. The ultimate goal is impossible to achieve by individual efforts, so the primary research goal of the thesis is to develop methods to systematically support the exploration of mechanisms that can explain the polarization of "private and unobservable opinions", and hence facilitate the approach to the ultimate goal. The primary goal can be decomposed into four sub-goals, and each of them intends to provide a helpful method, namely a formal polarization measurement that is consistent with the widely accepted understanding of polarization, a modeling method that can incorporate the "private and unobservable" feature of opinions in mechanisms, a modeling method that can help investigate the effect of non-default strategies (such as obfuscation), and a unifying framework that is expected to organize existing efforts and facilitate future studies. In particular, the framework will connect all the sub-goals by accommodating the other three methods as its own components. The four sub-goals have covered almost all aspects of a typical study of opinion polarization in computational sociology, and they collectively aim to provide systematical support for future studies.

1.3. RESEARCH APPROACH

As argued before, explaining the polarization of "private and unobservable opinions" is a complex task, which leads us to adopt a wide variety of research methods to accomplish the above-formulated goals. The four sub-goals lead to four studies, and the specific methods used for each study will be briefly described as follows.

For the study of the first sub-goal about polarization measurement, I mainly use axiomatic derivation (Esteban & Ray, 1994), which simply means that the measurement is derived from certain pre-given axioms and properties. Axiomatic derivation is a standard approach to developing new polarization measurements (e.g., Esteban & Ray, 1994; Wang & Tsui, 2000; Duclos et al., 2004) as it provides a rigorous theoretical foundation for mea-

surement development. In practice, the method would clearly state what kinds of axioms or properties I wish the measurement to possess and ensure that they will be possessed by the measurement. In most cases the axioms are used to characterize measurements directly without considering how the measurements will divide the data to be measured into groups, leading to the absence of the "group" notion in measuring polarization although it is crucial in understanding polarization. In my study, the method is used in a different way. By specifying a number of axioms and properties that have been deemed essential for an ideal polarization measurement, I first derive the grouping method that can construct the ideal measurement instead of directly deriving the measurement itself. The measurement is then constructed based on the grouping method in order to ensure that a proper group structure can be clearly identified in any data that the measurement will be applied to.

The major research method in both studies of the second and third sub-goal is agent-based modeling (ABM). ABM is a widely used modeling approach to investigating macroscopic phenomena (such as opinion polarization) that result from microscopic interactions (Bonabeau, 2002; Baumann, 2021). As the name indicates, ABM centers on agents, the autonomous entities that can represent particles, cells, organizations, countries, and in most opinion dynamics studies, people (Bonabeau, 2002; Ghorbani, 2013; Wilensky & Rand, 2015). Agents have their own attributes (such as opinions and behaviors), and can interact with each other or with the environment they live in according to some rules given by modelers. These microscopic interactions will update agents' attributes (and probably the environment as well), leading to possibly complex macroscopic phenomena that may reflect real-world dynamics (Bonabeau, 2002; Wilensky & Rand, 2015). ABM is especially popular when searching for opinion polarization mechanisms, as modelers can play with various (microscopic) conditions, such as interaction rules, initial distributions of heterogeneous agents, and landscapes of the environment (e.g., social network structure), in order to generate the macroscopic phenomenon of opinion polarization (e.g., Mäs & Flache, 2013; Mäs & Bischofberger, 2015). In fact, the modeling methods that will be established according to sub-goal 2 and 3 are specific implementations of ABM in the context of opinion polarization.

For the study of the final sub-goal, I follow an intuitive approach to building and validating a conceptual framework (see Coates et al. (2018) for an application). The building process includes reviewing the literature, categorizing relevant works, identifying relevant components, highlighting their relations, and finally illustrating the components and their relations, which will be the graphic representation of the framework. The validation process is twofold. First, I fit existing works into the framework by decomposing them into the framework components. Second, based on the framework I develop an agent-based model to explain a particular type of opinion polarization as a case study. The main aim is to evaluate the functionality of the framework: can the framework be used to organize existing works and facilitate future studies?

Apart from the above-mentioned methods that are specific to different research goals, the four studies driven by four sub-goals share the same "methodological structure", consisting of theory/ concept development, method development, and method implementation in case studies or illustrative examples of particular situations. The structure is similar to the so-called "methodological pyramid" in the sense that the previous step

1

of the structure is the basis of the next step, and the last step gives a demonstration of the method while (implicitly) evaluating the previous two steps (Diepenmaat, 1997). In the first study of polarization measurement, the measurement is applied to a two-dimensional synthetic data set to show the grouping process step by step; in the second study of the Action-Opinion Inference Model, the modeling method has been applied to various opinion-action relations (in forms of matrices) to investigate their effects on the patterns of public opinions; in the third study of obfuscation, I provide two illustrative examples to show how the modeling method works; in the last study of framework, as mentioned above, I fit existing works into the framework and develop a novel opinion polarization model using the framework.

1.4. OUTLINE AND CONTRIBUTIONS

The thesis comprises in total 6 chapters. Excluding Chapter 1 (Introduction) and 6 (Conclusions and Reflections), each of the remaining 4 chapters (hereafter referred to as "content chapters") is either a published paper or a paper to be submitted that aims to achieve one of the sub-goals. This paper-based format inevitably affects the unity of the entire thesis. To avoid possible confusion, the reader is advised to pay special attention to the inconsistency between chapters, especially in the usages of different terminologies (e.g., the term "action" is used in Chapter 3 and 4 as a synonym of "behavior" that is used in other chapters) and notations (e.g., opinion is denoted by r in Chapter 3 but by o in other chapters), although I have attempted to minimize the impact by making clarifications when necessary. In addition, spelling and citation styles in some chapters have been modified to be consistent with the rest of the thesis. Otherwise, each content chapter is identical to its published/ to be submitted version[5], whose information will be placed under the relevant chapter title in *Italic* font.

Being relatively concise, the thesis has a very straightforward structure: after the introduction given by Chapter 1, the order of the content chapters follows that of the sub-goals, namely Chapter $X+1$ attempts to fulfil sub-goal X ($X = 1,2,3,4$). In particular, Chapter 2-4 provide components of the framework proposed in Chapter 5. Finally, the thesis is concluded by Chapter 6.

The contribution of each remaining chapter is summarized as follows.

Chapter 2: Together Alone: A Group-based Polarization Measurement

Tang, T., Ghorbani, A., Squazzoni, F., & Chorus, C. G. (2021). Together alone: A group-based polarization measurement. Quality & Quantity.

DOI:10.1007/s11135-021-01271-y

This chapter presents a novel group-based polarization measurement that is derived from my newly proposed grouping method called "Equal Size Binary Grouping" (ESBG). ESBG would divide a given data set into two groups of equal sizes based on their similarities. The measurement is then defined as a function that increases with the between-group heterogeneity and decreases with the within-group heterogeneity. Apart from satisfying a list of long deemed desired properties such as being applicable to both

[5] A small number of obvious typos and spelling/ grammar mistakes have been corrected, while some reformulations have been made according to the feedback provided by the doctoral committee.

uni- and multi-dimensional data, the measurement is free from a number of theoretical and practical problems that have been troubling measurements that are derived from other grouping methods. Finally, the measurement is compared with the measurements of bimodality and bipolarization respectively, and its relation with the latter is further explained by the so-called "squeezing-and-moving" framework.

Chapter 3: Learning Opinions by Observing Actions: Simulation of Opinion Dynamics Using an Action-Opinion Inference Model

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).
DOI:10.18564/jasss.4020

This chapter provides a modeling method to highlight the "private and unobservable" feature of opinions in opinion dynamics (which may lead to opinion polarization) by presenting the Action-Opinion Inference (AOI) model. The central idea is that we can only learn the opinions of others by observing and inferring their actions, which is based on the realistic assumption that actions serve as the noisy signals of opinions. The AOI model introduces the so-called action-opinion matrix to encode the relations between opinions and actions, which not only guide an agent's choice of behavior based on her opinion but also direct the inference process when she has observed other agents' actions. The inference process would give the observer an estimation of the popularity of each opinion in her neighborhood, and she will adopt one of these opinions with the probability proportional to the estimation. Simulations of the model have revealed that the dynamics of opinions would be largely determined by the action-opinion relations, and with proper relations, the AOI model is capable of generating various patterns of opinions, including consensus, diversity, and polarization.

Chapter 4: Hiding Opinions by Minimizing Disclosed Information: An Obfuscation-based Opinion Dynamics Model

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

Based on the AOI model proposed in Chapter 3, this chapter presents a formal modeling method to investigate the role of obfuscation in opinion dynamics. Obfuscation refers to the opinion hiding strategy whereby people choose the actions that disclose the least information about their opinions. Compared to the "default strategy" of agents in the AOI model (i.e., choosing actions purely based on opinions), obfuscation represents a type of "deviate" strategy that arises naturally from the multiplicity in the action-opinion relations.

The modeling method – in the form of an "obfuscation-based opinion dynamics model" – integrates the obfuscation mechanism proposed by Chorus et al. (2021) with the AOI model, which enables us to distinctively model obfuscating and non-obfuscating agents, allowing for further analysis about how the share of (non-)obfuscating agents affects public opinion formation. By two illustrative examples from daily life and tales, I

show that the effect of obfuscation on opinion dynamics/ polarization is closely related to the action-opinion relations, although some qualitative conclusions can be derived.

Chapter 5: Modeling Opinion-Behavior Co-evolution for Explaining Opinion Polarization: a framework
Tang, T., Ghorbani, A., & Chorus, C. G. (2021). Modeling opinion-behavior co-evolution for explaining opinion polarization: A framework (to be submitted).

As the last content chapter, it provides a unifying framework – called MOBEP (**M**odeling **O**pinion-**B**ehavior co-evolution for **E**xplaining Opinion **P**olarization) – of the opinion-behavior co-evolution with the particular aim of explaining opinion polarization. On the basis of past attempts, I summarize that "behavior is the messenger of opinion", and identify five key components in modeling the co-evolution. The first four components, viz. "opinion-driven behavior change", "normative influence", "behavior-opinion inference", and "informational influence", constitute the so-called "behavior dynamics" which describes the situation where agents can only observe others' behaviors. The last component "direct opinion influence" alone constitutes the so-called "direct opinion dynamics", where agents can directly observe the opinions of others. In addition to these key components, each of the two dynamics is governed by its own social network or partner selection mechanism, while being connected with each other by an implementation component of schedule. The insights from Chapter 3 and 4 have been incorporated in the framework as part of the "opinion-driven behavior change" and "behavior-opinion inference" components, and the work of Chapter 2 contributes to the last implementation component of the framework– polarization measurement, which is needed by every opinion polarization study. To validate as well as demonstrate the framework, a number of existing models are decomposed according to the framework, and a case study of opinion polarization about mask wearing has been proposed, proving that as expected by sub-goal 4, the framework provides not only an overarching structure to organize existing efforts but also guidance for future studies.

Chapter 6: Conclusions and Reflections

The final chapter concludes the efforts made in this thesis to achieve the research goals formulated above, with a special focus on the relations between the proposed methods. In addition, several promising directions for future works are identified, followed by the societal relevance of the thesis. Finally, some notable reflections, including the outlooks of the field, are discussed.

The reader should note that the summarized contributions above only intend to provide a glimpse of these chapters, and many possibly interesting aspects, especially technical details, are not mentioned to avoid a lengthy introduction. Finally, I hope this not very lengthy introduction has been successful in initiating your interest in the rest of the thesis. If so, enjoy reading.

REFERENCES

[1] Abelson R. P. (1964). Mathematical models of the distribution of attitudes under controversy. In N. Frederiksen & H. Gulliksen (Eds.), Contributions to mathematical psychology (pp. 142-160). Rinehart Winston.

[2] Axelrod, R. (1997). The dissemination of culture: A model with local convergence and global polarization. Journal of Conflict Resolution, 41(2), 203-226.

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

[4] Baumann, F. T. W. (2021). Modeling opinion dynamics on networks: How social influence shapes the formation of consensus and polarization [Doctoral dissertation, Humboldt Universitaet zu Berlin]. edoc-Server: Open-Access-Publikationsserver der Humboldt-Universität. https://edoc.hu-berlin.de/handle/18452/23901

[5] Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. Proceedings of the National Academy of Sciences, 99(suppl 3), 7280-7287.

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

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

[8] Chitra, U., & Musco, C. (2020). Analyzing the impact of filter bubbles on social network polarization. In Proceedings of the 13th International Conference on Web Search and Data Mining (pp. 115-123).

[9] Chorus, C., Van Cranenburgh, S., Daniel, A. M., Sandorf, E. D., Sobhani, A., & Szép, T. (2021). Obfuscation maximization-based decision-making: Theory, methodology and first empirical evidence. Mathematical Social Sciences, 109, 28–44.

[10] 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).

[11] 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.

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

[13] Diepenmaat, H. B. (1997). Trinity: Model-based support for multi-actor problem solving [Doctoral dissertation, Universiteit van Amsterdam]. UvA-DARE (Digital Academic Repository). https://dare.uva.nl/search?identifier=ff508af9-2005-47c2-ad1f-55ce6407028f

[14] Duclos, J. Y., Esteban, J., & Ray, D. (2004). Polarization: concepts, measurement, estimation. Econometrica: Journal of the Econometric Society, 72(6), 1737-1772.

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

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

[17] 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).

[18] Ghorbani, A. (2013). Structuring socio-technical complexity: Modelling agent systems using institutional analysis [Doctoral dissertation, Technische Universiteit Delft]. TU Delft Research Repository. https://repository.tudelft.nl/islandora/object/uuid:ab80a728-e1da-4d6f-bb3a-9566180b2541

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

[20] Harteveld, E., (2021). Polarisation in the Netherlands: how divided are we? https://www.uva.nl/en/shared-content/faculteiten/en/faculteit-der-maatschappij-en-gedragswetenschappen/news/2021/01/elections-polarisation-in-the-netherlands-how-divided-are-we.html

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

[22] 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).

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

[24] 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.

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

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

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

[28] 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.

[29] McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. Annual Review of Sociology, 27(1), 415-444.

[30] 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.

[31] Pariser, E. (2011). The filter bubble: What the Internet is hiding from you. Penguin.

[32] 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.

[33] Spohr, D. (2017). Fake news and ideological polarization: Filter bubbles and selective exposure on social media. Business Information Review, 34(3), 150-160.

[34] Takács, K., Flache, A., & Mäs, M. (2016). Discrepancy and disliking do not induce negative opinion shifts. PloS ONE, 11(6), e0157948.

[35] Wang, Y. Q., & Tsui, K. Y. (2000). Polarization orderings and new classes of polarization indices. Journal of Public Economic Theory, 2(3), 349-363.

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

[37] 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

[38] 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.

[39] 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).

2

TOGETHER ALONE: A GROUP-BASED POLARIZATION MEASUREMENT

Tang, T., Ghorbani, A., Squazzoni, F., & Chorus, C. G. (2021). Together alone: A group-based polarization measurement. *Quality & Quantity*.
DOI:10.1007/s11135-021-01271-y

The growing polarization of our societies and economies has been extensively studied in various disciplines and is subject to public controversy. Yet, measuring polarization is hampered by the discrepancy between how polarization is conceptualized and measured. For instance, the notion of group, especially groups that are identified based on similarities between individuals, is key to conceptualizing polarization but is usually neglected when measuring polarization. To address the issue, this paper presents a new polarization measurement based on a grouping method called "Equal Size Binary Grouping" (ESBG) for both uni- and multi-dimensional discrete data, which satisfies a range of desired properties. Inspired by techniques of clustering, ESBG divides the population into two groups of equal sizes based on similarities between individuals, while overcoming certain theoretical and practical problems afflicting other grouping methods, such as discontinuity and contradiction of reasoning. Our new polarization measurement and the grouping method are illustrated by applying them to a two-dimensional synthetic data set. By means of a so-called "squeezing-and-moving" framework, we show that our measurement is closely related to bipolarization and could help stimulate further empirical research.

Keywords: Polarization measurement, Clustering, Bipolarization, Multi-dimensional polarization

2.1. INTRODUCTION

The vast and growing gap between liberals and conservatives, the prevalence of ideological extremes in public debate on social media, and the increasing income disparities within and across countries are shaking societies across the world. Researchers are required to conceptualize, define, and formally measure these various forms of polarization in a comprehensive manner. In the past few decades, polarization research has led to seminal contributions in a wide variety of sub-disciplines of social science, such as economics (e.g., Esteban & Ray, 1994; Alichi et al., 2016), political sciences (e.g., Abramowitz & Saunders, 2008; Hare & Poole, 2014) and sociology (e.g., Flache & Macy, 2011; Flache et al., 2017).

One insightful way to classify polarization is to distinguish whether there is a predefined group structure. For instance, the statement "the society is highly polarized in terms of political views" differs from the statement "the society is highly polarized in terms of political views *across ethnic groups*". Indeed, the latter requires a predefined group structure - in this case, based on ethnicity - whereas the former does not. This is a far-from-trivial issue because these two types of polarization reflect different social processes as well as different interpretations regarding all the possible cleavages cross-cutting our societies. A detailed explanation can be found in Section 2.2.

Like many other polarization studies (e.g., Duclos et al., 2004; Flache & Mäs, 2008; Anderson, 2011), here we are interested in measuring the former type of polarization, which is more difficult and controversial because of the absence of a predefined group structure. We argue that even if there is no group structure that can be predetermined via theoretical hypotheses, the notion of group is still crucial to achieve a more rigorous measurement of polarization. While pre-existing theories about social cleavages could be used to hypothesize the existence of precise group structures in our complex societies, developing consistent measurements of polarization that might help scholars discover group structures while scanning empirical data is still key to conceptualizing and understanding polarization.

Indeed, scholars have traditionally used groups to conceptualize and define polarization (e.g., Esteban & Ray, 2012; Danzell et al., 2019; Bauer, 2019). For example, to return to the above-mentioned example of a society being highly polarized in terms of political views, referring to groups is important: intuitively, a highly polarized society would consist of a small number of groups whose political views are very similar within the group, but very different between groups. The division of these groups would solely reflect each individual's political views without any reference to other factors, such as ethnicity or religion. In other words, instead of being imposed by "exogenous" factors, these groups would emerge endogenously from the variable(s) of interest (here: political views).

Unfortunately, little attention has been paid to such "endogenously emerging" group structures in polarization measurements, with the notion of group usually omitted (e.g., Flache & Mäs, 2008; Aleskerov & Oleynik, 2016) or penalized by various theoretical and practical problems (e.g., Esteban & Ray, 1994; Duclos et al., 2004). The usual difficulties of segmenting the social space of complex societies into group structures across various polarization dimensions would undermine the reliability of polarization measurements and so our understanding of the degree and extent of social polarization.

This paper aims to contribute to this field of research by proposing a novel way to

generate groups as the basis of a generic class of polarization measurements without predefined group structures. The method, called "Equal Size Binary Grouping" (ESBG), uses clustering techniques to assign people (data points) to two groups of equal sizes according to the variable(s) of interest. On the one hand, this method can help researchers to identify "endogenously emerging" group structures starting from data. On the other hand, this permits linking the concept of groups to the variable(s) of interest without losing relevant information, which is often involved in theory-driven, ex-ante group conceptualization. The group structure generated by ESBG overcomes various problems, such as discontinuity and contradiction of reasoning, leading to polarization measurements that satisfy a range of important properties that have long been deemed desired in the field (Esteban & Schneider, 2008; Gigliarano & Mosler, 2009). Furthermore, ESBG-based measurements are designed to measure both uni- and multi-dimensional polarization for discrete distributions. Although less frequently considered in literature, the latter has a great empirical value (see Section 2.4.1 for a more detailed discussion on this distinction).

The remainder of this paper is structured as follows: in Section 2.2, we provide a review of relevant literature concerning past attempts to conceptualize and measure polarization with or without predefined groups. We then propose a list of desired properties that an ideal polarization measurement without predefined groups should satisfy. In Section 2.3, we show that Equal Size Binary Grouping (ESBG) is a possible and promising approach to deriving the ideal polarization measurement (of a particular form) adhering to these properties while being free from the problems mentioned above. Inspired by clustering algorithms, Section 2.4 presents the procedure for implementing ESBG and constructing corresponding polarization measurements. An illustrative example using synthetic data is given in Section 2.5, followed by a series of discussions in Section 2.6 about the relation between the proposed polarization measurements and bipolarization measurements. The relation is further explained by the so-called "squeezing-and-moving" framework. Section 2.7 summarizes the study and draws conclusions.

2.2. BACKGROUND

2.2.1. POLARIZATION AND GROUPS

For decades, the concept of polarization has received ample attention in various fields, yet, without a consensual definition. For instance, in the field of international relations, polarization usually refers to "the degree of which antipathetic, non-overlapping subgroups are formed" (Hart, 1974), where these subgroups are defined according to the amity within each subgroup and the enmity between them. For example, the Allies and the Central Powers were two subgroups of nations during World War I. In economics, polarization is characterized as the "separation or distance across clustered groups in a distribution" (Esteban & Ray, 2012). Given their particular interest in income polarization, economists consider a society to be polarized when the population can be grouped into significantly sized groups of individuals having similar incomes within each group, which differ across groups (Esteban & Ray, 2012). In sociology, polarization in public opinion is conceptualized as "the degree to which the group can be separated into a small set of factions who are mutually antagonistic in opinion space and have maximal internal agreement" (Flache & Mäs, 2008), which mirrors the definitions in international relations

and economics.

These examples show that almost all definitions of polarization emphasize the notion of group, in the sense that members of the same group should be similar, and members of different groups should be dissimilar (in terms of the variable(s) of interest, such as income and opinion). Instead of the word "group", studies have used similar terms such as "clusters", "camps", "factions", or "subgroups". Regardless of the exact term being used, in all disciplines, groups, instead of individuals, are considered to be the crucial actor in conceptualizing polarization (Danzell et al., 2019).

In accordance with the development of polarization concepts, a growing number of polarization measurements have been formally proposed. A considerable portion of these measurements calculate polarization between groups that have been defined a priori[1] based on an external variable (hereafter referred to as *grouping variable*), a variable that is different from the variable of interest.

To clarify this: when one says "our society is polarized in terms of X across (or between) Y" then X is the variable of interest, and Y is the grouping variable. For instance, when using Gigliarano-Mosler (GM) index to measure the income and education polarization between East and West Germany, the grouping variable is the location of each individual (East or West Germany), while the variables of interest are income and education (Gigliarano & Mosler, 2009). Similar examples can be found in the measurement proposed by Zhang and Kanbur (ZK index) (2001), as well as Fusco and Silber (2014). These are sometimes called "social polarization measurements" (Fusco & Silber, 2014) or "socioeconomic polarization measurements" (Duclos & Taptué, 2015), as groups are usually defined by social characteristics such as race and religion. For the sake of clarity and simplicity, here we call **"polarization with exogenously imposed groups"** the one between groups that are explicitly defined by grouping variable(s) instead of variable(s) of interest, because in these cases the grouping variables are exogenous to the variables of interest. Note that this type of measurements and relevant studies have focused on the congruency between opinion and demographic attributes – a crucial factor affecting team performance – and thus gaining interest in organization and management literature (Phillips, 2003; Homan et al., 2007; Mäs et al., 2013).

However, in many other cases, it is more relevant to discuss polarization without exogenously imposed groups. Theoretically, polarization across particular socio-demographic strata (e.g., race, religion, ethnicity) is different from the polarization of the whole society. For instance, the opinion polarization of a society can be viewed as a result of opinion polarization across genders, races, locations, and countless other factors. Therefore, even if the degree of opinion polarization across one of these factors would be low, the society as a whole could still be highly polarized. Furthermore, there may also be practical objections to measuring polarization across exogenously imposed groups. Indeed, data of the grouping variables are not always available and in many cases, the only observation is the distribution of the variable(s) of interest. These arguments underline the importance of measuring polarization by defining groups in terms of the variable(s) of interest only.

Correspondingly, we call **"polarization with endogenously emerging groups"** the one where groups emerge based on the variable(s) of interest. Previous research has suggested two distinct lines of measurements of this type of polarization. The first line, started by

[1] i.e., the predefined groups as described in Section 2.1.

Wolfson (1994), measures polarization in terms of "the decline of the middle class (i.e., the group with a moderate value of the variable of interest)" (Foster & Wolfson, 2010). Therefore, the polarization measurement would be large whenever the middle class is negligible. The second line, founded by Esteban & Ray (1994), has the basic idea that a system is considered polarized if (i) the degree of heterogeneity within each group is low, (ii) the degree of heterogeneity across groups is high, and (iii) there is a small number of significantly sized groups (Esteban & Schneider, 2008).

Both lines are very popular, each with a large number of followers. The Wolfson's line is sometimes considered as the measurement of "bipolarization", which is conceptually different from the "polarization" measured by the Esteban & Ray's line (Deutsch et al., 2013). Furthermore, according to different sources of literature, bipolarization can be regarded as a category of polarization (Duclos & Taptué, 2015) or a concept that is distinct from polarization (Deutsch et al., 2013). We will interchangeably use the term "measurements in the Wolfson's line" and "bipolarization measurements". In this study, we primarily focus on the line originated by Esteban & Ray (1994). The relation between the two lines as well as our measurement will be further discussed in Section 2.6.

A common problem of the measurements in the Esteban-Ray's line concerns discontinuity. In the Esteban-Ray (ER) index (Esteban & Ray, 1994), polarization is measured by the effective antagonism, which is a function of identification within groups and alienation between groups. Here, groups are defined in a particularly sharp form whereby members of the same group must have exactly the same value of the variable of interest. To give an extreme example, people with an income of 1000 euro and 1000.01 euro are in two distinct groups. Esteban and Ray (1994) themselves have acknowledged the risk of sharp groups, namely the "discontinuity problem": there will be a jump in the polarization measurement if two close groups merge. It is difficult to justify such a jump, making these sharp groups theoretically implausible. The DER index (Duclos et al., 2004) and the Anderson's index (Anderson, 2011) can be viewed as the ER index of continuous variables and multi-variables respectively, and finding any group structure in these measurements is hardly feasible.

It is worth noting that a number of measurements are not covered by these two lines. The uncovered measurements may not involve the notion of endogenously emerging groups. For instance, in opinion dynamics literature, the FM index calculates the variance of the pairwise distance for all pairs of individuals (Flache & Mäs, 2008; Flache & Macy, 2011). Therefore, the notion of group is not included. A more recent example is the Schweighofer-Schweitzer-Garcia (SSG) index (Schweighofer et al., 2020), which is a function of the sum of squared pairwise differences. For multi-dimensional polarization (where there is more than one variable of interest), Aleskerov and Oleynik (2016) consider a multi-dimensional variable as a vector, and define "center of mass" as the weighted average of all vectors. Polarization is then measured by the weighted sum of the distances between each vector and the center of mass.

Table 2.1 provides an overview of the polarization concepts and measurements mentioned above[2]. It suggests that although the notion of group is crucial in defining and conceptualizing polarization, there has been no rigorous way to formalize it in order

[2]The measurements in the Wolfson's line will be introduced in Section 2.6. The WT index refers to the measurement from Wang & Tsui (2000).

to measure polarization with endogenously emerging groups. We acknowledge that all measurements mentioned here were developed for particular research questions, and hence the absence of group structures would be acceptable to achieve simple polarization measurements. However, we believe that with the intention to better understand and measure polarization, there should be an appropriate polarization measurement that clearly tells us what the group structure is, and how to measure polarization based on it.

Table 2.1: Summary of the concepts and measurements of polarization

Types of polarization		Corresponding measurements
with exogenously imposed groups		ZK index, GM index
with endogenously emerging groups	ER's line	ER index, DER index
	Wolfson's line	Wolfson's index, WT index
	Others	FM index, SSG index

2.2.2. PROPERTIES OF POLARIZATION MEASUREMENT

In order to tackle this problem, we propose a generic class of polarization measurements based on a novel method to define groups according to the variable(s) of interest. The method, called "Equal Size Binary Grouping" (ESBG), divides the population into two groups of equal sizes on the basis of similarities within each group and between different groups. We will show that a polarization measurement generated by this method, subject to certain requirements in the constructing procedure, satisfies various properties that have long been deemed desired in the field, including:

- **Continuity**: the measurement is a continuous function.
- **Dimensionality**: the measurement can be applied to both uni- or multi-dimensional discrete data.
- **Monotonicity**: the measurement decreases with within-group heterogeneity and increases with between-group heterogeneity
- **Maximum and Minimum**: the measurement is maximized when the population is equally divided into two maximally dissimilar groups, and members in the same group have the same value of the variable of interest. The measurement is minimized when everyone has the same value of the variable of interest.
- **Normalization**: The measurement should be in the range of 0 to 1.

The properties of continuity and normalization are important not only because of their omnipresence in literature (e.g., Esteban & Ray, 1994; Chakravarty & Majumder, 2001; Gigliarano & Mosler, 2009), but also because a continuous and normalized polarization measurement is much easier to analyze than a discontinuous and non-normalized one.

The property of dimensionality echoes the growing interest in multi-dimensional polarization (Aleskerov & Oleynik, 2016). We will further discuss this in Section 2.4.1.

The importance of the monotonicity property is widely acknowledged in polarization studies with exogenously imposed (Zhang & Kanbur, 2001; Gigliarano & Mosler, 2009) as well as endogenously emerging groups (Esteban & Ray, 1994). Herein, while within-group heterogeneity refers to the heterogeneity or dissimilarity of members in the same group, between-group heterogeneity refers to the heterogeneity or dissimilarity between members of different groups. Different polarization measurements may use different expressions for these two variables. In many measurements of polarization with exogenously imposed groups, heterogeneity is represented by inequality (Zhang & Kanbur, 2001; Gigliarano & Mosler, 2009). In the ER index, given that each group only contains people with the same value of the variable of interest, the within-group heterogeneity is always zero and the between-group heterogeneity is simply the absolute difference between groups.

It is worth noting that in studies of polarization with endogenously emerging groups, while polarization level typically decreases with within-group heterogeneity (Esteban & Schneider, 2008), there is no clear conclusion about the relation between polarization level and between-group heterogeneity. The only thing that has been confirmed is that the degree of between-group heterogeneity must be high in a highly polarized system (Esteban & Ray, 1994; Esteban & Schneider, 2008). Such a relatively vague description, which may be due to the lack of properly defined groups (see Section 2.3), breaks the symmetry and brings difficulty in polarization analysis. Ideally, we would like to propose polarization measurements that not only decrease with *within-group* heterogeneity but also increase with *between-group* heterogeneity.

The importance of the maximum and minimum property has been highlighted in previous research (Gigliarano & Mosler, 2009; Flache & Macy, 2011; Fusco & Silber, 2014; Bauer, 2019). Particularly, there is hardly any polarization measurement that violates the maximum property regardless of how polarization is conceptualized. The minimum property indicates that a polarization measurement should be minimized at perfect equality, and originates from the so-called "normalization axiom" (Chakravarty & Majumder, 2001). For instance, in studies of opinion polarization, the minimum condition refers to the state of consensus where everyone has the same opinion (Flache & Mäs, 2008; Flache & Macy, 2011; Schweighofer et al., 2020).

In addition, we emphasize that an ideal polarization measurement should also satisfy a number of axioms that have been used in constructing measurements (Esteban & Ray, 1994), and are subject to some practical constraints, which will be further discussed in Section 2.3.4.

2.2.3. AN ALTERNATIVE APPROACH TO MEASURING POLARIZATION

While all measurements previously discussed have tried to capture the overall picture of polarization by one single expression, there are alternative approaches that measure polarization in different aspects with respective indices, especially in sociological research. For instance, DiMaggio et al. (1996) suggest four distinct dimensions – dispersion measured by variance, bimodality measured by kurtosis, constraint (association between different dimensions of the variable of interest) measured by Cronbach's alpha, and consolidation (association between variable of interest and exogenously imposed groups) measured by "differences in groups' means over time" (McCright & Dunlap, 2011). Bramson et al. (2016,

2017) decompose polarization into nine "senses", namely spread, dispersion, coverage, regionalization, community fracturing, (endogenously emerging) group distinctness, group divergence, group consensus, and group size parity. These dimensions and senses are largely overlapping and highly correlated.

As regards political polarization in the United States, Boxell et al. (2017) consider eight indices, each capturing a particular part of political polarization, such as: partisan affect polarization, ideological affect polarization, and partisan sorting. While these indices are mostly related to DiMaggio's dimensions and Bramson's senses, there are specificities that reflect the particular case of American politics. The point here is that each individual index alone is unable to reflect the whole picture, and this may lead to conflicting assessments. To fill this gap, Boxell et al. (2017) not only applied all eight indices to the data set, but also constructed an overall index of polarization based on the average of all indices. The advantage of this approach is twofold. First, as most aspects already have their own pre-existing measurements, scholars can easily apply them to their data sets, saving the effort of constructing a new measurement. Second, this approach displays more information than the single-expression approach, allowing scholars to discover trends or draw conclusions for each aspect.

However, there are drawbacks in this approach (DiMaggio et al., 1996; Bramson et al., 2016, 2017; Boxell et al., 2017). First, knowing how many aspects and which aspects are sufficient to capture polarization is hard. Therefore, choosing the optimal set of aspects can be difficult. Moreover, depending on different scenarios, certain aspects could be particularly salient while others would not. For instance, among DiMaggio's four dimensions, Baldassarri & Bearman (2007) only use dispersion and bimodality, deciding to ignore others. Second, different aspects can be correlated and overlapping, thus making it difficult to design the overall index especially for quantitative research that aims at comparability, replicability, and cumulativeness.

Finally, it is worth noting that even in these alternative approaches, the concept of group, whether exogenously imposed or endogenously emerging, is still key to measuring polarization. For endogenously emerging groups, Bramson et al. (2017) define groups "directly from the histogram" of the distribution plot. Further analysis of this type of grouping methods (in the context of bimodality) and its comparison with our method (i.e., ESBG) can be found in Section 2.4.5.

2.3. DERIVATION OF EQUAL SIZE BINARY GROUPING

When you have eliminated the impossible, whatever remains, however improbable, must be the truth. (Sherlock Holmes)

The aim of this section is to justify ESBG as an appropriate grouping method for constructing ideal polarization measurements. To achieve this aim, after clarifying the notations (Section 2.3.1), we will show that ESBG is a possible solution to the problems afflicting other grouping methods (Section 2.3.2): the grouping method without any constraints suffers from the discontinuity problem (G_0), and the grouping method only constrained by a fixed number of groups contradicts Esteban and Ray's reasoning (Esteban & Ray, 1994) (G_1). Furthermore, in Section 2.3.3, we will explain how ESBG takes into account the roles of the missing variables, namely the number and size of groups,

by providing some examples. Finally, in Section 2.3.4, we will test if the ESBG-based polarization measurement satisfies the axioms proposed by Esteban & Ray (1994).

2.3.1. NOTATIONS

We first present the following notations that will be used throughout the rest of the paper. Suppose we are interested in a discrete system (i.e., data set) $X \equiv \{\mathrm{x}_1, ..., \mathrm{x}_N\}$ consisting of N data points. A data point $\mathrm{x}_i = (x_{i,1}, ..., x_{i,D})$ $(i = 1, ..., N)$ is described by its variables $x_{i,d}$ $(d = 1, ..., D)$, where D is the dimension of the system. A grouping method $G : X \to C$ partitions the system X into K non-overlapping groups $C = \{C_1, ..., C_K\}$. The size of a group C_k is denoted by s_k, representing the number of data points in C_k. The within-group heterogeneity of a group C_k is denoted by w_k, and the between-group heterogeneity of a pair of groups C_i and C_j $(i \neq j,\ i, j = 1, 2, , , K)$ is denoted by $b_{i,j}$. In Section 2.4, we will further discuss how to calculate w_k and $b_{i,j}$.

At the end of Section 2.2, we have listed a range of desired properties that an ideal polarization measurement should adhere to. Assume now that we already have completed the task of partitioning the data set into groups. Then the polarization measurement should be a function of at least the following two factors: within-group heterogeneity and between-group heterogeneity. Intuitively, the number of groups (K) and the size of each group $(S = \{s_1, ..., s_K\} \in \mathbb{R}^K)$ may also affect the polarization level. Therefore, such a polarization measurement should have the following form:

$$P(X) = f(W, B, K, S) \tag{2.1}$$

where $W \in \mathbb{R}^+$ and $B \in \mathbb{R}^+$ are indices for within-group heterogeneity and between-group heterogeneity of the entire data set respectively. As the desired properties suggest, P should be decreasing with W and increasing with B. Following Gigliarano & Mosler (2009) where choices of W and B are mostly related to the weighted sum of each group's characteristics, we further assume that the two variables should take the following forms:

$$W = \phi\left(\sum_{k=1}^{K} \alpha_k w_k\right) \tag{2.2}$$

$$B = \psi\left(\sum_{i<j} \beta_{i,j} b_{i,j}\right) \tag{2.3}$$

where ϕ and ψ should be strictly increasing and continuous. $\alpha_k > 0$ and $\beta_{i,j} > 0$ are real number coefficients, representing the weights or importance of corresponding variables. There are good reasons for using such linear expressions $(\sum_{k=1}^{K} \alpha_k w_k$ and $\sum_{i<j} \beta_{i,j} b_{i,j})$ as inputs of ϕ and ψ. As we will see in Section 2.3.2, 2.3.3, and 2.3.4, the linearity will significantly simplify our analysis about the properties of P, by making it possible to directly obtain the changes in W and B during certain dynamical processes. These changes may be intuitive and are not of our main interest here, but not giving specific forms or using other expressions of W and B might make the formal derivation of the outcome tedious and difficult, if still possible. For example, in Figure 2.1, there are three groups at I_1, I_2, and I_3 $(I_1 < I_2 < I_3)$. If I_1 and I_2 move to each other for the same distance, we intuitively anticipate that $B = \psi(b_{1,2}, b_{2,3}, b_{1,3})$ should decrease, but it is not easy to prove: it is unclear if B will decrease as $b_{1,3}$ decreases but $b_{2,3}$ increases. Nonlinear forms

of ψ, such as product, may require extensive efforts to confirm the result, while expression (2.3) can solve it easily through simple calculation (see Section 2.3.2). This will become clearer in the rest of the section thanks to some further examples. Given the benefit of linearity, and the lack of advantage of nonlinear expressions, we choose equation (2.2) and (2.3) for the rest of the paper.

Following the linearity in W and B, in this section we further take the following assumption: $b_{i,j}$ should be the squared distance between the centers (or mean values) of C_i and C_j. Similarly, w_k should be the average squared distance between members of C_k and the center (mean value) of C_k. We do not aim to rule out other forms of $b_{i,j}$ and w_k, but this assumption will significantly simplify our analysis in Section 2.3.2. For example, in Figure 2.2, there are two groups: C_1 that contains people at 1 and 5, and C_2 that contains people at 11. If people at 1 and 5 move towards each other with the same distance, we can easily show that $b_{1,2}$ stays fixed with this assumption.

2.3.2. SEARCHING FOR GROUPING METHODS

The lack of a proper grouping method is the root for the absence of well-defined "endogenously emerging" groups in polarization measurements. In general, a grouping method is a series of steps that separate the data set into a finite number of (non-overlapping) groups, where members of the same group should be similar and members from different groups should be dissimilar according to some criteria. Additionally, multiple constraints – including the number and size of groups – can be applied to a grouping method based on prior knowledge or specific requirements. In this subsection, we consider three types of grouping methods: method without any constraint, method with a fixed number of groups, and method with both fixed number and size of groups. Conceptually, the three types represent all possible grouping methods. We will show that a particular grouping method of the last type, called "Equal Size Binary Grouping" (ESBG), which divides the data set into two equally sized groups, should be a possible solution to problems such as discontinuity and contradiction of reasoning if we want to construct an ideal polarization measurement that (i) is in the form of equation (2.1), (2.2), and (2.3), and (ii) adheres to the desired properties.

A common requirement for endogenously emerging groups is that they should be formed on the basis of (dis)similarities between individuals (i.e., data points), so that each group is homogeneous internally but different from other groups. Let us assume that all the grouping methods discussed in this subsection satisfy this requirement. This implies that each of them is able to classify data points that are sufficiently similar into the same group and classify the data points that are sufficiently dissimilar into different groups. We will leave the question "how to perform these grouping methods to ensure that they satisfy this requirement" to Section 2.4, where more technical details will be provided.

GROUPING METHOD WITHOUT ANY CONSTRAINT

Suppose there is a grouping method G_0 whose *only* task is to divide the system into groups. Therefore, there is no constraint on G_0 besides the requirement mentioned above, and the number and size of the groups are determined to make the members of the same group similar and members of different groups dissimilar.

To understand why G_0 is not a proper grouping method for the polarization measure-

ment $P(X)$, consider the uni-dimensional example given in Figure 2.1, modified from Esteban & Ray (1994). In Figure 2.1, initially (at $t = 0$), half of the population is equally distributed between level $I = I_1$ and $I = I_2$ (I is the variable of interest), and the other half of the population is at level $I = I_3$. Suppose $0 < I_1 < I_2 < I_3$, $I_3 - I_2 \geq I_2 - I_1$, and the three levels are sufficiently different such that G_0 will produce three non-overlapping groups C_1, C_2, and C_3, containing data points at I_1, I_2 and I_3 respectively. Therefore, $w_k = 0$ ($k = 1, 2, 3$). Now, consider that both C_1 and C_2 move towards each other synchronously with the same speed until merging. $X(t)$ is the system at time t. During the process, there must be a moment $t = t^*$ when C_1 moves to I_1^* ($I_1 < I_1^* < I_2$) and C_2 moves to I_2^* ($I_1^* < I_2^* < I_2$) [3], and G_0 **starts** to recognize C_1 and C_2 as one group, denoted by C_4. The transition moment t^* fully depends on G_0 if the moving speed is given. When $t < t^*$, the between-group heterogeneity $B = \psi(\beta_{1,2} b_{1,2} + \beta_{2,3} b_{2,3} + \beta_{1,3} b_{1,3})$ is decreasing (with an intuitive condition that $\beta_{1,3} = \beta_{2,3}$) [4]. Due to the fact that W, K, and S are constant, the decrease in B implies that $P(X)$ decreases with t when $t < t^*$. When $t > t^*$, there will be only two groups C_4 and C_3, and $W = \phi(\alpha_4 w_4 + \alpha_3 w_3)$ decreases with t as w_4 is decreasing, while other factors stay constant; therefore $P(X)$ is increasing. To conclude, $P(X(t))$, as a function of t, is decreasing when $t < t^*$ and is increasing when $t > t^*$.

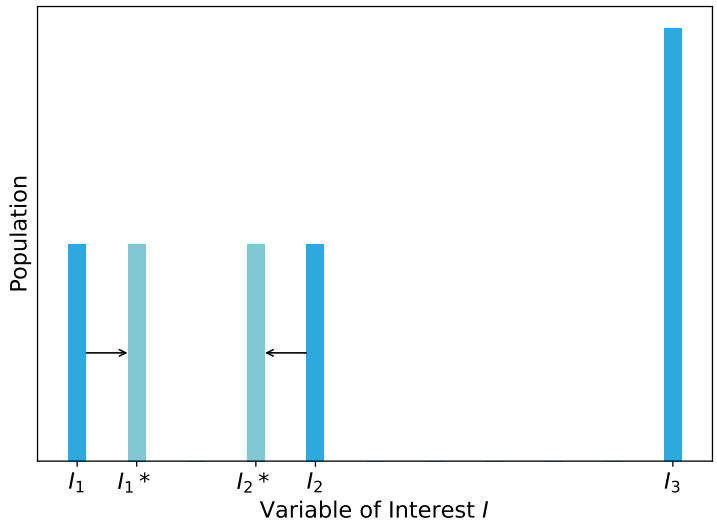


Figure 2.1: Diagram to illustrate the failure of G_0.

Assume now that we have a new grouping method G_0^{II}, which is slightly different from

[3] If $I_1^* = I_2^*$, G_0 is the same as the ER index. See Section 2.2.1.

[4] It is clear that the difference between C_1 and C_2 (i.e., $b_{1,2}$) shrinks, so we only need to prove that $\beta_{2,3} b_{2,3} + \beta_{1,3} b_{1,3}$ is not increasing. As it is intuitive to have $\beta_{1,3} = \beta_{2,3}$, the task is reduced to prove $b_{1,3} + b_{2,3}$ is not increasing. Suppose C_1 or C_2 has traveled a distance of Δ ($0 < \Delta < (I_1^* - I_1)$) at time t' ($t < t' < t^*$) since t (C_1 and C_2 always travel the same distance), then $[b_{1,3}(t') + b_{2,3}(t')] - [b_{1,3}(t) + b_{2,3}(t)] = [(d_{1,3}(t) - \Delta)^2 + (d_{2,3}(t) + \Delta)^2] - [(d_{1,3}(t))^2 + (d_{2,3}(t)^2] = 2(d_{2,3}(t) + \Delta - d_{1,3}(t))\Delta < 0$, where $d_{i,j}(t)$ is the distance between I_i and I_j at t. Therefore $b_{1,3} + b_{2,3}$ is indeed decreasing. See Section 2.3.1 for the reason of choosing $b_{i,j} = (d_{i,j})^2$.

G_0 in the sense that the transition moment for G_0^{II} is $t^{**} > t^*$. Denote the polarization measurement of X using G_0 as $P(X|G_0)$ and using G_0^{II} as $P(X|G_0^{II})$. When $t < t^*$ or $t > t^{**}$, G_0 and G_0^{II} are of no difference and hence $P(X|G_0) = P(X|G_0^{II})$. $P(X|G_0) = P(X|G_0^{II})$ when $t < t^*$ implies that $\lim_{t\uparrow t^{**}} P(X(t)|G_0^{II}) < P(X(t = t^{**})|G_0)$ if we assume both $P(X|G_0)$ and $P(X|G_o^{II})$ are continuous in t, as $P(X|G_0)$ is increasing during $t^* < t < t^{**}$, and $P(X|G_0^{II})$ is decreasing during the same period. Meanwhile, $P(X|G_0) = P(X|G_0^{II})$ when $t > t^{**}$ implies that $\lim_{t\downarrow t^{**}} P(X(t)|G_0^{II}) = P(X(t = t^{**})|G_0)$; therefore $\lim_{t\uparrow t^{**}} P(X(t)|G_0^{II}) < \lim_{t\downarrow t^{**}} P(X(t)|G_0^{II})$, which directly proves that $P(X(t)|G_0^{II})$ is discontinuous at $t = t^{**}$. Given that there are countless transition moments generated by countless grouping methods without any constraint, we can conclude that $P(X(t))$ is a discontinuous function of t. Note that not only does such discontinuity exist in our example; it is likely to occur whenever two (or even more than two) groups merge.

Indeed, this problem of G_0 is the same as the discontinuity problem observed in the ER index (see Section 2.2.1). Besides being counter-intuitive, this discontinuity will cause various problems. For example, the sudden jump of the polarization level at the transition moment is hardly justifiable. If such discontinuity is accepted, one can dramatically increase or decrease the polarization level of the same data set by simply constructing a slightly different transition moment.

GROUPING METHOD WITH FIXED NUMBER OF GROUPS

To solve the discontinuity problem in G_0, we impose a constraint on G_0: the number of groups is fixed to $K = 2$. We denote this grouping method as G_1. The task of G_1 is to divide the systems into two groups such that the two groups are maximally different, but members in the same group are maximally similar. To show that G_1 overcomes the discontinuity problem, we also apply G_1 to $X(t = 0)$ in Figure 2.1. Since $I_3 - I_2 > I_2 - I_1$, individuals at I_1 and I_2 are classified into one group, say, C_4, and individuals at I_3, as before, constitute the other group C_3. Now the dynamic process described in Figure 2.1 only decreases w_4, while w_3, B, and S are not affected. Therefore $P(X(t)|G_1)$ increases continuously throughout the process, that is, it does not suffer from discontinuity.

When I_2 is closer to I_3 than to I_1, is the method still discontinuity-free? Again, consider that both individuals at I_1 and I_2 move towards each other with the same speed simultaneously. Initially, G_1 will define two groups: C_1 containing everyone at I_1 and C_5 containing everyone at I_2 or I_3. There will also be a transition moment $t = t^*$ when G_1 starts to consider individuals at I_1^* and I_2^* as one group C_4. The question is if $P(X|G_1)$ is discontinuous. Through simple analysis, we know that $P(X|G_1)$, again, decreases with t when $t < t^*$ and increases with t when $t > t^*$. However, it is intuitive to see that no matter which G_1 we choose, the transition moment t^* is always the moment when $I_3 - I_2 = I_2 - I_1$. Otherwise, the group structure will violate the basic requirement mentioned in the beginning of this section. Therefore, the method is free from discontinuity.

Although providing a solution to discontinuity, G_1 has its own problem. Consider another example modified from Esteban & Ray (1994). As shown in Figure 2.2, almost all individuals are placed equally at $I = 1$ and $I = 5$, while only a sufficiently small number of individuals are at $I = 11$. G_1 will put individuals at 1 and 5 in the same group, say C_1, leaving those at 11 in another group C_2. Now, consider all individuals in C_1 merge at $I = 3$. The merge reduces w_1 to 0, while all other factors remain unchanged. Consequently, the

polarization level should go up. However, according to Esteban & Ray (1994), due to the relatively small size of C_2, the initial polarization mostly comes from the dissimilarity between the individuals at 1 and 5, which is eliminated after the merge, so the polarization level should go down. This contradiction discourages using G_1 for constructing $P(X)$. Note that choosing another value for K not only lacks a strong theoretical justification but also is unable to solve the discontinuity problem in Figure 2.1.

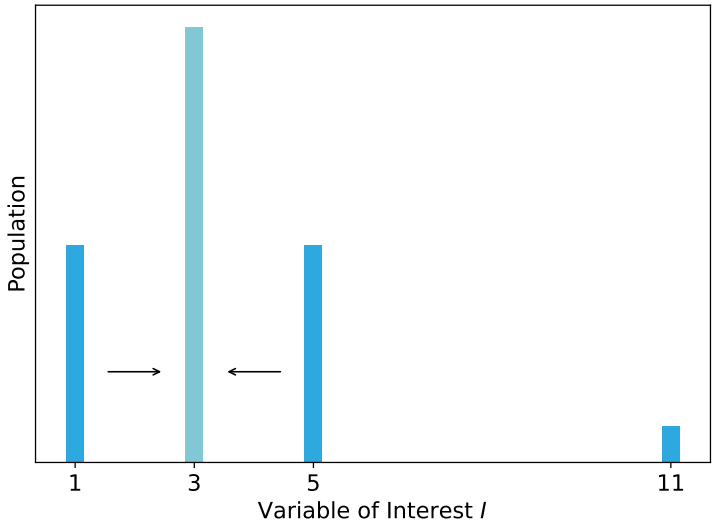


Figure 2.2: Diagram to illustrate the failure of G_1.

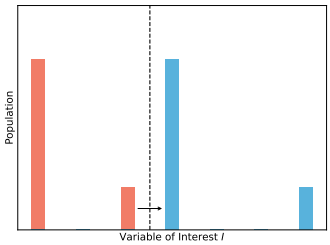
GROUPING METHOD WITH FIXED NUMBER AND SIZE OF GROUPS

From the analysis above, we know that the problem in G_1 comes from group size. To solve this problem, we impose another constraint on G_1: the size of each group must be the same. We call it **Equal Size Binary Grouping** (**ESBG**), whose task is to split the system into two equally sized groups, while maximizing between-group heterogeneity and/or minimizing within-group heterogeneity. For the sake of simplicity, we only discuss systems whose size is an even number. We will discuss how to implement this method in Section 2.4.

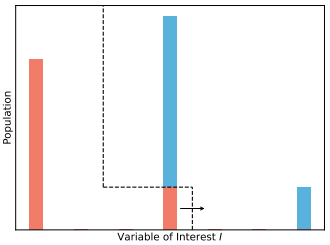
First, the discontinuity problem in Figure 2.1 can be completely solved by replacing G_0 with ESBG as there will be no transition moment during the process. For a more general case, consider the dynamic process described in Figure 2.3, in which ESBG initially divides the uni-dimensional system into two groups: C_1 in red and C_2 in blue (Figure 2.3(a)). Without loss of generality, we suppose that a portion of C_1 move towards C_2, and will stop after passing the closest member of C_2 (Figure 2.3(c)). Note that the colors in the figure only indicate the **initial** group memberships.

We can see that compared to G_0, moving sufficiently close to individuals of another group can no longer trigger a transition of group membership under ESBG (Figure 2.3(a)). Only when the moving red individuals pass the blue individuals near the group boundary,

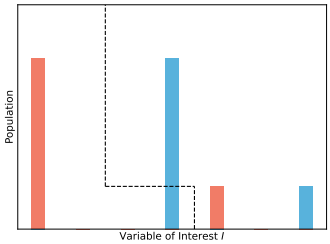
there will be a transition of group membership as the moving individuals, previously members of C_1, will be now identified as members of C_2 (Figure 2.3(c)). However, we can take an alternative look at this situation. The identity of an individual is purely determined by its value of the variable of interest; therefore among the individuals in the middle in Figure 2.3(b), ESBG cannot tell which individual just moved here from the left (red), and which individual is native (blue). Therefore, the dynamics from Figure 2.3(b) to Figure 2.3(c) can be equivalently interpreted as the dynamics from Figure 2.3(b) to Figure 2.3(d), that is, a portion of native blue individuals move to the right, and no membership transition happens during the whole process. Due to the arbitrariness of this example, we can conclude that the discontinuity problem caused by group membership transition can be solved by ESBG.



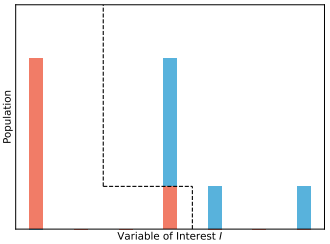
(a) Beginning of the process



(b) Fusion



(c) End of the process



(d) An alternative view of (c)

Figure 2.3: Illustration of the continuity of ESBG. Colors indicate the initial membership of each individual; arrows represent the moving direction; and the dashed line is the current group boundary.

Finally, we show that ESBG has the potential to solve the problem found in G_1. In the example given in Figure 2.2, ESBG will define groups differently from G_1: all individuals at 1 and a small number of individuals at 5 will form C_1, and the rest of the population will form C_2. During the merging process in Figure 2.2, we can confirm that B decreases given a sufficiently small population at 11[5]. Meanwhile, w_1 decreases and w_2 increases, so it

[5]Note that this confirmation needs a sufficiently small size of individuals at 11. Denote the members of C_1 at 1 as C_1^1, the members of C_1 at 5 as C_1^5, the members of C_2 at 5 as C_2^5, and the members of C_2 at 11 as C_2^{11}. During the merging process, the heterogeneity between C_1^1 and both C_2^5 and C_2^{11} decreases, while the heterogeneity between C_1^5 and C_2^5 is fixed at 0. Meanwhile, the heterogeneity between C_2^{11} and C_1^5 increases. However, if the size of C_2^{11} is small enough (the size of C_1^5 is smaller than C_2^{11}), we can conclude that B decreases.

is unclear whether W increases or not. However, it gives us room to design expressions for W, B, and $f(W,B)$ in order to solve the problem, which is much better than G_1 where polarization will definitely increase. For example, Section 2.5.3 shows a particular implementation of f which should be able to solve the problem of G_1 (see Table 2.2).

To summarize, both G_0 and G_1 are not qualified as grouping methods for constructing the ideal polarization measurement because of discontinuity and contradiction of reasoning, while ESBG should be a possible solution.

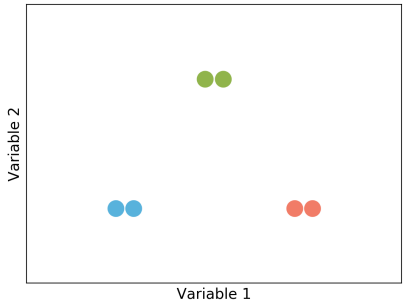
2.3.3. ESBG AND THE MISSING VARIABLES

By using ESBG, the expression $P = f(W,B,K,S)$ reduces to $P = f(W,B)$. However, removing K and S does not mean the measurement fails to include the effects of these two variables. Indeed, the role of K and S are inherited by W and B. In practice, $S = \{s_1, ..., s_K\}$ is represented by the **relative group sizes** (RS), which measures "how equally populated the groups are" (Gigliarano & Mosler, 2009). Large RS implies that the group sizes are similar, and small RS implies unequal distribution of group sizes.

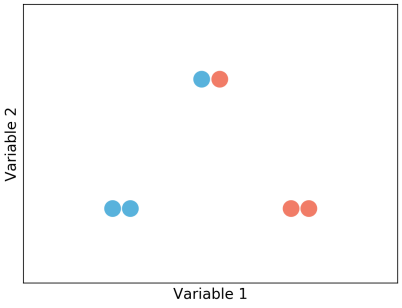
Figure 2.4 and 2.5 provide vivid examples in a two-dimensional space. In Figure 2.4(a), a constraint-free grouping method (i.e., G_0) divides the data set X_1 into three groups, each containing two identical individuals. The distance (i.e., heterogeneity) between each group is assumed to be the same. Given the same data set, ESBG will divide X_1 into two groups, which means one of the three groups defined by G_0 (in Figure 2.4(a), the green group) will be equally separated and taken by the remaining groups (Figure 2.4(b)). Now, consider another data set X_2 where G_0 divides it into two groups (i.e., the blue and the red groups), each containing three identical individuals (Figure 2.4(c)). ESBG will make the same division (Figure 2.4(d)) as G_0. It is not difficult to find out that (a) and (c) have the same W, B and RS[6]. Therefore, the only difference between X_1 and X_2 under G_0 (i.e., (a) and (c)) is the number of groups, K. However, when using ESBG to measure polarization (i.e., (b) and (d)), both data sets have the same $K = 2$. Comparing (b) and (d), we find that the data set with a larger K (i.e., X_1) under G_0 will have a larger W but a smaller B under ESBG, indicating that X_1 is less polarized than X_2 not only under G_0 but also under ESBG (intuitively a polarization measurement is negatively related to K when $K \geq 2$). To conclude, the effect of K under G_0 is replaced by the effect of W and B under ESBG.

Figure 2.5 shows how ESBG converts the effect of RS to the effect of W and B. In Figure 2.5(a), G_0 divides the data set X_3 into two groups, each containing two or four identical individuals (i.e., the blue and red group). Meanwhile, ESBG will divide X_3 into two groups, each containing three members as shown in (b). Figure 2.5(c) and (d) show that both G_0 and ESBG divide another data set X_4 into two groups, each containing two identical individuals. Assume the distances between the two groups in (a) and (c) are the same, then the only difference between X_3 and X_4 under G_0 is RS. Without any calculation, it is clear that X_3 has a smaller RS than X_4, indicating that X_4 is more polarized (because intuitively a polarization measurement is positively related to RS). Comparing (b) and (d) (where RS no longer matters), X_3 has a larger W and a smaller B, indicating that X_3 is less polarized than X_4 under ESBG, in line with the prediction by G_0. Therefore, the effect of S (via RS) under G_0 is replaced by W and B under ESBG.

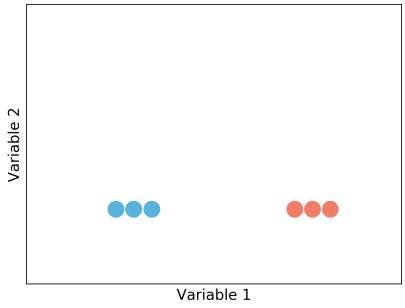
[6] Both (a) and (c) have reached the maximum of RS (see Gigliarano & Mosler (2009) for details). A condition for (a) and (c) to have the same B is $\sum_{i<j}^{K} \beta_{i,j} = C, \forall K > 0$, where C is a constant.



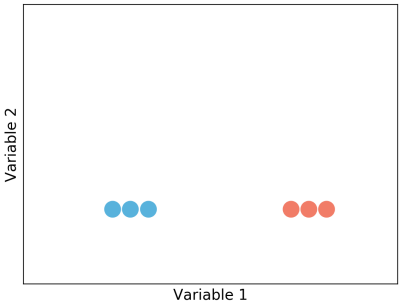
(a) Apply G_0 to X_1



(b) Apply ESBG to X_1

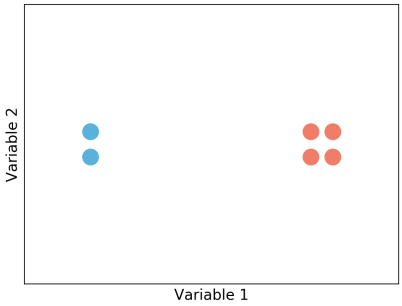


(c) Apply G_0 to X_2

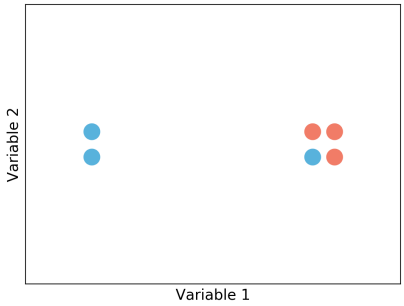


(d) Apply ESBG to X_2

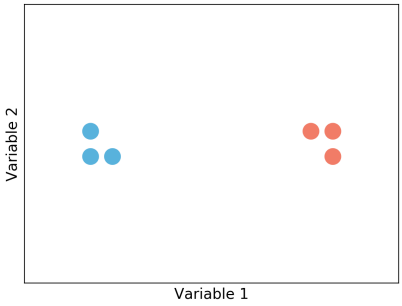
Figure 2.4: Applying G_0 and ESBG to data set X_1 and X_2 respectively. In (a),(c), and (d), points of the same color are not only in the same group but also identical, while in (b), color only represents group membership. To avoid overlap and improve readability, positions of the data points have been adjusted.



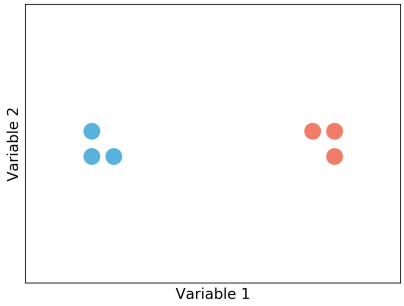
(a) Apply G_0 to X_3



(b) Apply ESBG to X_3



(c) Apply G_0 to X_4



(d) Apply ESBG to X_4

Figure 2.5: Applying G_0 and ESBG to data set X_3 and X_4. In (a),(c), and (d), points of the same color are not only in the same group but also identical, and the distance between points with the same color does not represent the difference between them but is introduced to improve readability. In (b), color only represents group membership.

2.3.4. POLARIZATION AXIOMS BY ESTEBAN & RAY

In this subsection, we test whether an ESBG-based polarization measurement, even without a particular expression, satisfies the axioms proposed by Esteban & Ray (1994). For the sake of simplicity, we reduce equation (2.2) and (2.3) to $W = \phi(w_1 + w_2)$, and $B = \psi(b_{1,2})$, which will be further justified in Section 2.4.3.

Axiom 1 (Figure 2.6)

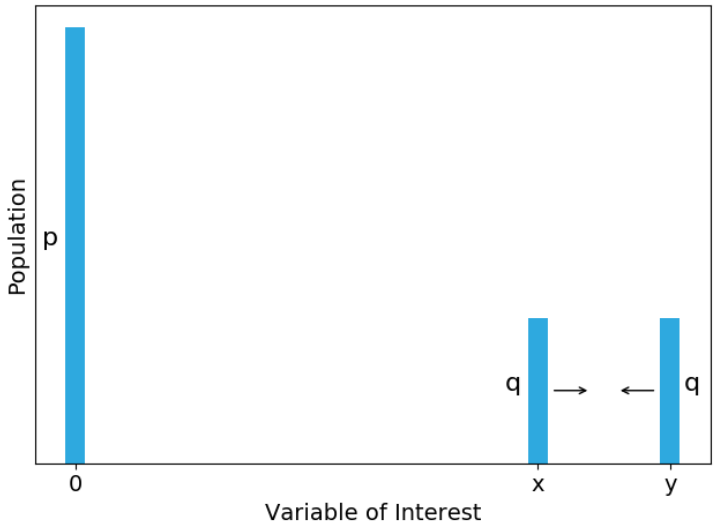


Figure 2.6: Esteban-Ray's Axiom 1:
Data: $p, q \gg 0$, $p > q$, $0 < x < y$.
Statement: Fix $p > 0$ and $x > 0$. There exist $\epsilon > 0$ and $\mu > 0$ such that if $d(x, y) < \epsilon$ (d is the distance function) and $q < \mu p$, the joining of the two q masses at their mid-point, $(x + y)/2$, increases polarization. Note: This statement, as well as Axiom 2 and 3, are directly taken and modified from Esteban & Ray (1994).

To justify this statement, assume that μ is small enough such that $2q < p$ (i.e., $\mu < 1/2$). Therefore, under ESBG, the two q masses are always in the same group, say, C_2. A part of the p mass will also be in C_2, and the rest of the p mass will be in the other group C_1. Given that the merge does not affect the center of C_2, B is not affected. Meanwhile w_1 is obviously not affected, but w_2 decreases, which will increase $f(W, B)$.

The condition $d(x, y) < \epsilon$, in the original paper (Esteban & Ray, 1994), was proposed to ensure that the two q masses were sufficiently close. Under ESBG, this condition is no longer needed.

Axiom 2 (Figure 2.7)

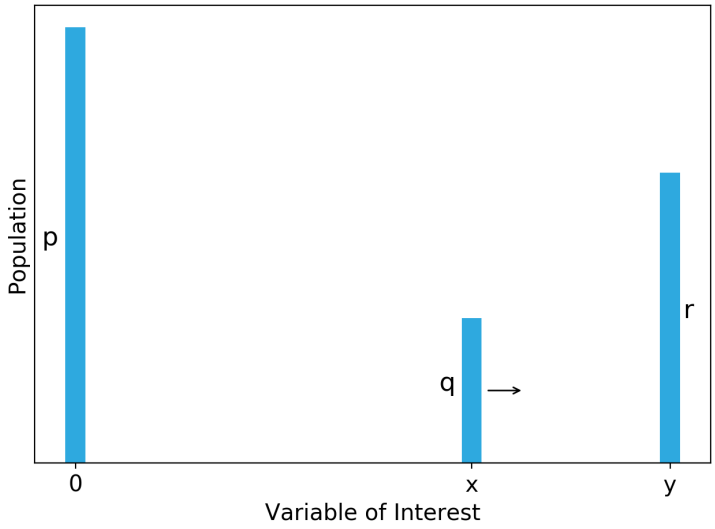


Figure 2.7: Esteban-Ray's Axiom 2:
Data: $(p,q,r) \gg 0$, $p > r$, $x > |y - x|$.
Statement: There exists $\epsilon > 0$ such that if the population mass q is moved to the right (towards r) by an amount not exceeding ϵ, polarization goes up.

If $p > (p+q+r)/2$, q mass, r mass, and a part of the p mass will be in the same group C_2, while the rest of the p mass will be the other group C_1. After the move, w_1 is not affected and B goes up. If w_2 decreases, we have P increasing as the axiom requires. If w_2 increases, it does not increase as much as B does (given that B and W are on the same scale): on the one hand, the move of the q mass will decrease the heterogeneity between the q and r mass, which deteriorates w_2; on the other hand, the move increases the heterogeneity between the q mass and everyone in the p mass. However, there are more members of C_1 than members of C_2 in the p mass, implying that this move should affect B much more than w_2. Therefore, polarization should go up after the move given a properly designed measurement. Here we provide a simple intuition rather than a formal proof.

If $p = (p+q+r)/2$, the move will decrease w_2 and increase B, thus $f(W,B)$ will increase.

If $r < p < (p+q+r)/2$, more than half of the q mass will be in C_2, while the rest of the q mass will be in the other group C_1. The move of the q mass increases w_1 but reduces w_2, which should lead to a decrease in $w_1 + w_2$ because more than half of the q mass is in C_2. Due to the same reason, the move will increase B. As a result, $f(W,B)$ will go up.

Axiom 3 (Figure 2.8)

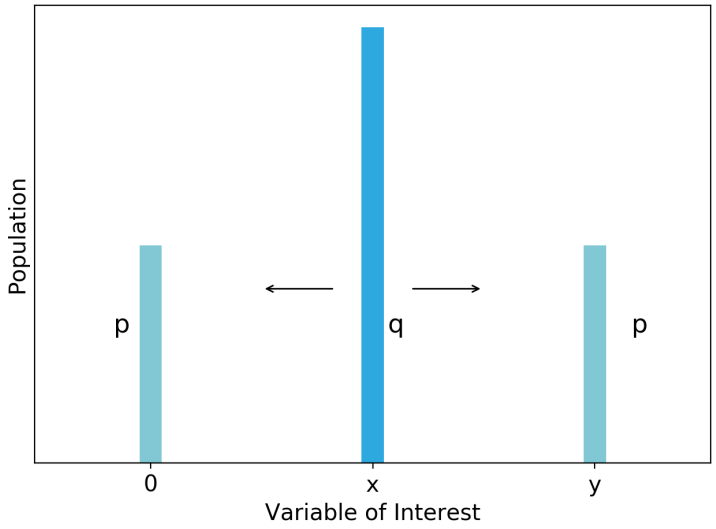


Figure 2.8: Esteban-Ray's Axiom 3:
Data: $(p, q) \gg 0$, $x = y - x \equiv d$.
Statement: Any new distribution formed by shifting population mass from the central mass q equally to the two lateral masses p, each d units of distance away, must increase polarization.

Before the split, $W = B = 0$. After the split, $W = 0$ and $B > 0$, and hence the polarization measurement goes up.

2.4. IMPLEMENTING GROUPING METHOD AND CONSTRUCTING POLARIZATION MEASUREMENT

As argued above, the task of ESBG is to split the data set into two equally sized groups in such a way that members from different groups are very different, and members within each group are very similar. In Section 2.3, we did not discuss how ESBG achieves this task or how to implement ESBG, but took it for granted. In this section, using ideas of clustering techniques, we propose an implementation protocol for ESBG, especially in a multi-dimensional space, based on which the ESBG-based polarization measurement will be constructed. In addition, we compare the ESBG-based measurement with bimodality measurements as they share similar expressions.

2.4.1. DIMENSIONALITY

If we are only interested in polarization of a uni-dimensional data set, implementing ESBG is as simple as dividing the data set by the median value. In this subsection, we will stress the importance of multi-dimensional polarization, justifying the necessity of implementing ESBG in multi-dimensional spaces.

Multi-dimensional polarization is not the simple aggregation of uni-dimensional polarization from different dimensions. Therefore, measuring multiple uni-dimensional polarization cannot tell how polarized the whole system is. Following the example given by Ross (1920), consider a society with half white men and half black men. Therefore, the society is ethnically polarized; meanwhile, the society consists of half employees and half employers, so it is also polarized in social classes. If all white men are employed by black men, the society is polarized as a whole with half white employees and half black employers. However, if half white/ black men are employers and half white/ black men are employees, the society is actually split into four groups: white employer, white employee, black employer, and black employee. Therefore, the society is polarized in all dimensions, but is less polarized as a whole. This also creates problems of micro vs. macro level measurements, as suggested by research on group segregation in labor markets (e.g., Takács et al., 2018).

The majority of polarization measurements are designed to measure uni-dimensional data only. Given the necessity of measuring multi-dimensional polarization, implementing ESBG for both uni- and multi-dimensional data sets is of paramount importance.

2.4.2. CLUSTERING

Clustering is one of the most important topics in data analysis and machine learning, which has been extensively studied due to its broad functionality (Jain et al., 1999; Xu & Wunsch, 2009). In short, clustering is "the task of partitioning the data set into groups" (Müller & Guido, 2016), and members in the same group "should display similar properties based on some criteria" (Xu & Wunsch, 2009). A twin concept of clustering is called supervised classification that depends on a set of pre-classified/ pre-labeled data. From the pre-classified data (also called "training data"), a supervised classification technique learns how to define groups, and then divides unlabeled data into groups (Jain et al., 1999). Unlike supervised classification, clustering deals with unlabeled data only, which means groups "are obtained solely from the (unlabeled) data" (Jain et al., 1999). This feature

naturally reminds us of the "endogenously emerging groups" that are solely derived from the variable(s) of interest, indicating that clustering is fundamentally similar to the task of defining endogenously emerging groups. Thanks to the development in the field, there is a vast collection of efficient and reliable clustering algorithms (see Jain et al. (1999), Baraldi & Blonda (1999a, b), Xu & Wunsch (2009) for reviews), which will pave the way for the implementation of ESBG.

A typical clustering process usually consists of the following steps (Jain & Dubes, 1988; Jain et al., 1999; Xu & Wunsch, 2009):

Feature selection and/or extraction: It is a necessary preprocessing step for clustering. Because not all features (in our context, dimensions) are "equally relevant" for clustering (Aggarwal, 2014), for the sake of efficiency, feature selection chooses the most relevant and effective set of features for defining groups (Jain et al., 1999). In addition, feature extraction transforms original features into new forms that are more salient.

Definition of a proximity measurement: As argued above, data points are clustered into groups according to how "close" they are to each other. To implement clustering, we need to formally define a proximity measurement. The term "proximity" is the counterpart of "homogeneity" in the context of polarization. Therefore, measuring proximity in clustering echoes measuring within- and between-group homogeneity/ heterogeneity in EBSG.

Grouping/optimization: This is the main step of clustering. Given the proximity measurement, the grouping step is "an optimization problem with a specific criterion function" (Xu & Wunsch, 2009), and the criterion is closely related to the proximity measurement.

Validation: This step assesses the output produced by previous steps depending on some optimal criteria (Jain et al., 1999).

2.4.3. IMPLEMENTING ESBG

Based on the steps of a clustering process, a formal ESBG process should include the following steps:

Preprocessing: This step mirrors the feature selection and/or extraction step in clustering. A common concern about multi-dimensional polarization is the incommensurability of dimensions. For instance, when measuring the two-dimensional polarization of education and income, it is difficult to defend why an x year difference in education and a y euro difference in income are equally important (x and y are arbitrary positive numbers). Furthermore, some less relevant dimensions might harm the efficiency of ESBG. The preprocessing step should help to solve these issues by techniques such as dimension reduction and rescaling (for details, see Aaberge & Brandolini (2015)). After the preprocessing, these dimension-related problems should no longer exist in the processed data.

Definition of a heterogeneity measurement: In this step, we need to design a heterogeneity measurement appropriate to our data. Just like proximity (Jain et al., 1999),

heterogeneity can be measured by a distance function – for example, the Euclidean distance – of pairs of data points. Once the heterogeneity measurement is chosen, by denoting the two groups as C_1 and C_2, the expressions of heterogeneity within each group (w_1 and w_2) and between groups ($b_{1,2}$) can be determined. The within-group heterogeneity W and between-group heterogeneity B should be calculated according to the following equations:

$$W = w_1 + w_2 \tag{2.4}$$

and

$$B = b_{1,2} \tag{2.5}$$

Equation (2.4) and (2.5) are the reduced forms of equation (2.2) and (2.3) respectively. The omission of the parameters α_1 and α_2 in equation (2.4) is due to the equity of group sizes. At the same time, we take the simplest possible expression of ϕ: $\phi(w_1 + w_2) = w_1 + w_2$. For the expression of B, since there are now only two groups, the overall between-group heterogeneity B is the same as the heterogeneity between C_1 and C_2, i.e., $b_{1,2}$.

Grouping: Given the expression of W and B, ESBG is translated into an optimization problem with the aim of maximizing B and/or minimizing W, subject to the constraint that the group number must be 2, and the sizes of the two groups must be the same.

Validation: The validation process in ESBG is almost the same as in clustering, but with an additional exam on the number and size of groups.

In practice, it is easy to choose a well developed clustering algorithm as the basis of ESBG. In Section 2.5, we will develop the implementation of ESBG based on the famous K-means clustering algorithm.

2.4.4. CONSTRUCTING A POLARIZATION MEASUREMENT

Given the endogenously emerging groups C_1 and C_2 defined by ESBG, a polarization measurement should take the following form:

$$P(X) = f(W, B) = f(w_1 + w_2, b_{1,2}) \tag{2.6}$$

since we have used $W = w_1 + w_2$ and $B = b_{1,2}$ (see Section 2.4.3).

When designing the expression for the measurement, it is important to ensure that all the desired properties for $P(X) = f(W, B)$ listed in Section 2.2.2 have been taken into account. These properties are formally summarized as follows:

Continuity: $P = f(W, B)$ is a continuous function of both W and B.
Dimensionality: $P : \mathbb{R}^D \to \mathbb{R}$. $D = 1$ or $D \geq 2$.
Monotonicity: $P = f(W, B)$ is strictly decreasing with W and strictly increasing with B.
Maximum: P is maximized when $W = 0$ and B is maximized.
Minimum: P is minimized when $W = B = 0$.
Normalization: For all $X \in \mathbb{R}^D$ ($D \geq 1$), $0 \leq P(X) \leq 1$.

Combining the maximum property and the normalization property, we have $f(W = 0, B = B_{max}) = 1$, where B_{max} is the maximum between-group heterogeneity. However, in practice, determining the value of B_{max} can be troublesome. Suppose we define B_{max} as the maximal pairwise distance in a data set X_1, and then we design an expression of $f(W,B)$, say f_1, such that $f_1(W(X_1) = 0, B(X_1) = B_{max}) = 1$. Then for another data set X_2 whose between-group heterogeneity $B(X_2)$ is larger than B_{max}, we will obtain $f_1(W(X_2) = 0, B(X_2)) > 1$, violating the normalization property. To solve the issue, we introduce the normalizing parameter $\delta > 0$, which should be greater than or equal to the maximum possible heterogeneity in all the data sets of interest. Formally, if we want to compare the polarization level of X_1, X_2,..., and X_M, then $\delta \geq max_m\, max_{x_i,x_j \in X_m} h(x_i, x_j)$ for all $m = 1, ..., M$, and h is the heterogeneity function. Note that in the rare case when all $h(x_i, x_j)$ are zero, δ can take an arbitrary positive value as it no longer matters. The normalizing parameter should then replace B_{max}, that is, $f(W = 0, B = \delta) = 1$. Once the value of δ is determined, it should stay constant for all data sets that are going to be compared. There is a variety of ways to determine the parameter. For example, in a recent opinion dynamics study where the data points are all in the range of -1 and $+1$, Schweighofer et al. (2020) use the "maximally possible distance" between two points in the opinion space as the normalizing parameter. This means in a D dimensional Euclidean space, their normalizing parameter is $\sqrt{4D}$.

Finally, we provide a particular form of $P = f(W,B)$ that exhibits the desired properties. It is worth noting that equation (2.7) is by no means the only possible form of $f(W,B)$.

$$f(W,B) = \frac{1}{\delta}\, g\left(\frac{B}{W+1}\right) \tag{2.7}$$

where g is a continuous and strictly increasing function with $g(0) = 0$ and $g(1) = 1$. It is easy to prove that this form satisfies the property of continuity, dimensionality, monotonicity, maximum, and normalization. It is also obvious that when $W = B = 0$, $f(W,B)$ defined in equation (2.7) is minimized to 0. One may argue that $B = 0$ and $W \neq 0$ can also lead to $P = 0$. However, our definition of groups implies that when $B = 0$, W must also be 0. Therefore $W = B = 0$ is a sufficient and necessary minimization condition.

Figure 2.9 summarizes the procedure of implementing ESBG and constructing polarization measurement based on ESBG. After preprocessing, the raw data are transformed to the "trouble-free" processed data. Subsequently, by defining the heterogeneity measurement, as well as the expressions of within-group heterogeneity W and between-group heterogeneity B, we divide the processed data into two groups of equal sizes. The grouping result needs to be validated. This concludes the procedure of implementing ESBG. To construct the polarization measurement, besides W and B, we further need to design the expression of $f(W,B)$ and choose an appropriate normalizing parameter δ. By applying the measurement to the groups, we can finally obtain the polarization level of the data.

2.4.5. RELATION WITH BIMODALITY MEASUREMENTS

The expression given by equation (2.7) resembles a number of bimodality measurements such as Ashman's D (Ashman et al., 1994; Forchheimer et al., 2015) and the bimodal separation (Zhang et al., 2003), whose main ideas lie in the assumption that the data set

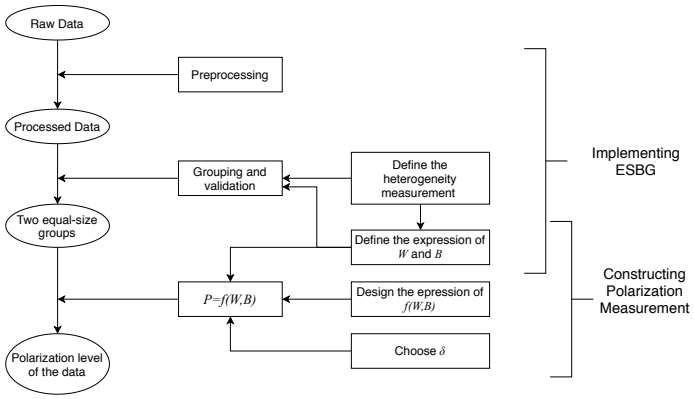


Figure 2.9: Procedure of implementing ESBG and constructing polarization measurement.

X is generated or can be described by some bimodal Gaussian mixture. The density of such a mixture is (Ashman et al., 1994):

$$p(X) = \pi_1 p(X, \mu_1, \delta_1^2) + \pi_2 p(X, \mu_2, \delta_2^2) \tag{2.8}$$

where π_g, μ_g, and δ_g^2 (not to be confused with the normalizing parameter δ) are the fraction, mean, and variance of a Gaussian distribution g ($g = 1, 2$). Given these parameters, Ashman's D is expressed as (Forchheimer et al., 2015):

$$D \propto \frac{|\mu_1 - \mu_2|}{\sqrt{\delta_1^2 + \delta_2^2}} \tag{2.9}$$

and the bimodal separation is (Zhang et al., 2003):

$$BS \propto \frac{|\mu_1 - \mu_2|}{\delta_1 + \delta_2} \tag{2.10}$$

The polarization measurement in equation (2.7) and the above-mentioned bimodality measurements both rely on the ratio of between-group heterogeneity to within-group heterogeneity if we consider each Gaussian distribution as a group. To use D and BS, one usually needs to fit two Gaussian distributions to the data set by some technique (e.g., the KMM algorithm (Ashman et al., 1994)), which is in analogy with ESBG. With all these commonalities, it is fair to conclude that the ESBG-based polarization measurements systematically echo the bimodality measurements as they all require a bi-division of the data set and use the heterogeneity between and within the divisions.

These similarities reflect the conceptual closeness between polarization and bimodality. DiMaggio et al. (1996) regarded bimodality as one of the four key dimensions of polarization. Bramson et al. (2017) argued that bimodality takes into account at least three "senses" of polarization, including community fragmentation ("the degree to which the population can be broken into sub populations" (Bramson et al., 2016, 2017)), distinctness between groups, and distance between groups (see Section 2.2.3). Bimodality is also

claimed to be an indicator (Knapp, 2007) or a feature (Bramson et al., 2017) of polarization. In fact, bimodality (not necessarily D and BS introduced here) has been used as a (partial) measurement of political polarization (e.g., Baldassarri & Bearman, 2007; Kim & Baek, 2021) and polarization has been used as an interchangeable (yet problematic[7]) term for bimodality (e.g., Hegselmann & Krause, 2002).

Despite these similarities, these two types of measurements are fundamentally different. Polarization and bimodality are distinct concepts. A bimodal distribution is usually polarized, but a distribution with zero bimodality (e.g., a unimodal distribution) may not be of zero polarization. As pointed out by Fiorina & Abrams (2008), bimodality is a necessary but hardly sufficient condition for a large degree of polarization. According to Bramson et al. (2017), bimodality does not implicitly invoke the sense of group size parity, which refers to the idea that a system is more polarized if groups are of equal sizes. This is reflected by the KMM algorithm where the size of each distribution is not relevant. Meanwhile, the ESBG-based measurement, as argued in Section 2.3.3, includes the effect of group size parity by imposing the equal size constraint.

2.5. AN ILLUSTRATIVE EXAMPLE: EQUAL SIZE BINARY GROUPING BASED ON K-MEANS CLUSTERING AND CORRESPONDING POLARIZATION MEASUREMENT

In this section, we provide an illustrative example of applying ESBG to a synthetic multi-dimensional data set and then constructing a polarization measurement based on the groups defined by ESBG.

2.5.1. K-MEANS CLUSTERING

The implementation of ESBG in this example will utilize one of the most well-known and widely-used clustering algorithms called **K-means clustering** (Forgy, 1965; MacQueen, 1967; Xu & Wunsch, 2009). Despite its ease of implementation, K-means clustering is an ideal choice for ESBG because the number of groups K needs to be determined a priori. To produce two groups, we simply set $K = 2$, and the only remaining problem is to ensure the sizes of the groups are equal.

In short, K-means clustering attempts to find a number (K) of centroids (sometimes called group/cluster centers), each representing a group containing the data points around the centroid (Müller & Guido, 2016). Formally, the algorithm divides the system by minimizing the following distortion function (Bishop, 2006):

$$J = \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} ||x_i - \mu_k||^2 \tag{2.11}$$

where $r_{ik} \in \{0, 1\}$ is a binary indicator: $r_{ik} = 1$ if x_i is classified in C_k, and $r_{ik} = 0$ otherwise. The distortion function J in Equation (2.11) is the sum of squared distances between each data point x_i and its centroid μ_k. The K-means algorithm chooses the optimal $\{r_{ik}\}$ and $\{\mu_k\}$ to minimize J by using an iterative procedure based on the EM algorithm: given

[7] According to Bramson et al. (2017).

randomly chosen initial conditions, during each iteration, first we fix μ and minimize J with respect to r_{ik} (step E); we then fix μ and minimize J with respect to r_{ik} (step M). The iteration is repeated until convergence (Bishop, 2006).

From equation (2.11), we can see the proximity measurement used in K-means clustering is the (squared) Euclidean distance. Meanwhile, the distortion function J is closely related to within-group heterogeneity (see Section 2.5.2). Therefore, from the view of grouping method, we can say that the K-means clustering algorithm defines groups by minimizing the within-group heterogeneity of the data set.

2.5.2. ESBG BASED ON K-MEANS CLUSTERING

In this subsection, we show how to implement ESBG step by step on the basis of the K-means clustering algorithm.

Preprocessing: We use a synthetic two-dimensional data set X^* containing two blobs of 100 and 200 data points respectively [8]. Assuming none of the dimension-related problems (see Section 2.4.3) exists, no preprocessing is needed for this particular case.

Definition of heterogeneity measurement: Following the K-means clustering algorithm, we use the squared Euclidean distance as the heterogeneity measurement. Within-group heterogeneity W can thus be defined as:

$$W = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{2} r_{ik}||x_i - \mu_k||^2 = \frac{1}{N}\sum_{k=1}^{2}\sum_{x_i \in C_k} ||x_i - \mu_k||^2 \tag{2.12}$$

which is the average squared Euclidean distance between each data point and its corresponding centroid, that is, $W = J/N$ with a predefined value of $K = 2$. Between-group heterogeneity B, following the same fashion, is defined as the squared distance between the centroids:

$$B = ||\mu_1 - \mu_2||^2 \tag{2.13}$$

The motivation for choosing J/N instead of J as the measurement of W is to make W and B on the same scale. Otherwise, we could expect $W \gg B$ in most cases, making P extremely small. In addition, if W and B are not on the same scale, for example $W = J$, it will be difficult to defend the second axiom of Esteban & Ray (1994) (see Section 2.3.4).

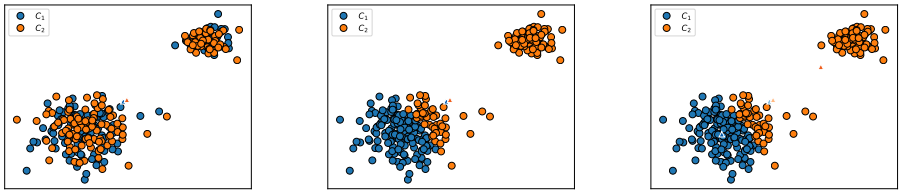
Grouping: Now the task of implementing ESBG is turned into an optimization problem:

$$\begin{aligned} \min_{\{r_{ik}\},\{\mu_k\}} \quad & W = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1,2} r_{ik}||x_i - \mu_k||^2 \\ \text{s.t.} \quad & \sum_{i=1}^{N} r_{ik} = N/2, \qquad \forall k = 1,2 \end{aligned} \tag{2.14}$$

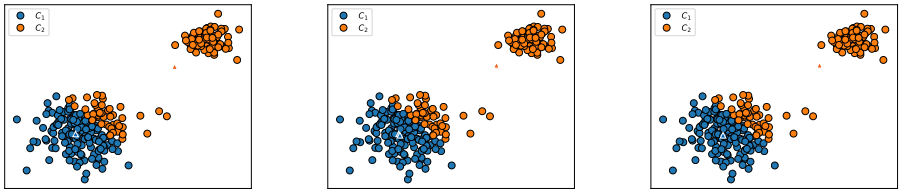
Following Bishop (2006), we use the EM iteration to solve this optimization problem. The M step is the same as that in K-means, and the E step aims to minimize W with fixed

[8] The data set X^* is generated using the *make_blobs* function from *sklearn* Python module: X,y=make_blobs(n_samples=[100,200], cluster_std=[.4,.8], centers=[[4,0],[0,-4]]). Then apply the MinMaxScaler from the same module and we obtain X^*.

μ_1 and μ_2, while constrained by the condition of equal group sizes. The basic idea is to calculate the squared distance between each data point and both centroids, respectively. For data point x_i, denote the absolute difference between its squared distances to both centroids as Δ_i. First we assign each data point to the closer centroid to generate "temporary" groups. Then, until both groups have the same size, we select a member repeatedly to move it from the larger group to the smaller group. The selected member should be the one with the smallest Δ_i. A similar idea can be found from the elki[9] project. The process is illustrated in Figure 2.10. The outcome is two equal-size groups with members distributed around two centroids (Figure 2.11(a)). To make a comparison, we apply K-means clustering to the same data set in Figure 2.11(b).



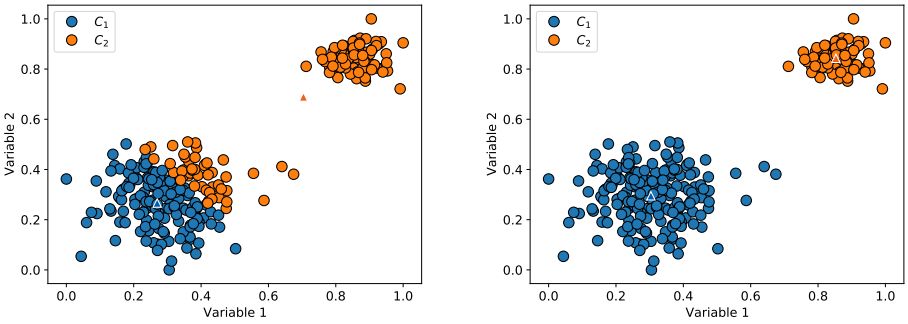
(a) Iteration 0: Initialization (b) Iteration 1: E Step (c) Iteration 1: M Step



(d) Iteration 2: E Step (e) Iteration 2: M Step (f) Iteration 3: E Step

Figure 2.10: Illustration of the K-means-based ESBG using X^*. The centroids are shown by the triangles of similar colors of their corresponding groups. The triangles of lighter colors represent the centroids in the previous iteration. (a): Initially, the data set is randomly and equally divided into two groups C_1 and C_2, and the centroids of both groups are computed as the average of their group members. (b): In the E step of Iteration 1, each data point is assigned to the group whose centroid is nearer, while keeping the size of each group equal. (c): In the M step of Iteration 1, the centroid of each group is re-computed according to the new group structure updated in the last E step. (d)-(f): Successive iterations. The change in the positions of centroids from (d) to (e) is relatively small and can be observed when taking a closer look. The system has reached convergence since (f).

[9] https://elki-project.github.io/tutorial/same-size_k_means



(a) ESBG (b) K-means

Figure 2.11: The results of applying (a) ESBG or (b) K-means clustering ($K = 2$) to X^*. Each centroid is shown by the triangle of the similar colour of its corresponding group. Note: (a) is the same as Figure 2.10 (f).

Validation: We first check if the outcome of the grouping step contains two groups, and if their sizes are the same. Then, we check whether the outcome is optimal, in other words, if W is minimized. A primary step may include checking if swapping memberships of data points can decrease W, and if each centroid is the mean of its members. In addition, a number of validation methods, criteria, and indices are available to formally justify the clustering result (Xu & Wunsch, 2009).

2.5.3. CORRESPONDING POLARIZATION MEASUREMENT

According to equation (2.7), we choose $f^*(W,B) = \frac{1}{\delta}(\frac{B}{W+1})$ as our polarization measurement (i.e., we choose $g(x) = x$). Given this expression, we have $W = 0.05784060054$, and $B = 0.367362082$ for our synthetic data set X^*. Setting $\delta = 2$ (given that the maximum possible squared Euclidean distance in X^* is smaller than $\sqrt{2}$), the polarization level of X^* is then $f^* = 0.173637731$.

As suggested in Section 2.3.2, we use f^* to examine whether ESBG can solve the problem of G_1. In Table 2.2, we have summarized the within-group heterogeneity before (W) and after (W^m) the merge, the between-group heterogeneity before (B) and after (B^m) the merge, and the polarization level before ($f^* = f^*(W,B)$) and after ($f^{*m} = f^*(W^m,B^m)$) the merge, under different initial population distributions at 1, 5, and 11. By setting $\delta = 100$, we can see that as long as the population at 11 is no larger than 8 (recall that in the original example, it is required that the population at 11 is sufficiently small), the merge will reduce the polarization level due to the significant decrease in B and relatively small increase in W.

2.6. RELATION WITH BIPOLARIZATION MEASUREMENTS

As argued in Section 2.2, there are two notable lines of polarization measurements: the Wolfson's line (i.e., bipolarization measurement), which captures the decline of the middle class, and the Esteban & Ray's line, which focuses on how individuals are clustered in groups. It is clear that our ESBG-based measurement is in the Esteban & Ray's line as

Table 2.2: Measuring polarization in the system described in Figure 2.2

Pop at 1	Pop at 5	Pop at 11	W/W^m	B/B^m	f^*/f^{*m}
9	9	2	3.6000/ 5.1200	23.0400/ 2.5600	0.0501/ 0.0042
8	8	4	5.6000/ 7.6800	31.3600/ 10.2400	0.0475/ 0.0118
7	7	6	6.0000/ 7.6800	40.9600/ 23.0400	0.0585/ 0.0265
6	6	8	4.8000/ 5.1200	51.8400/ 40.9600	0.0894/ 0.0669
5	5	10	2.0000/ 0.0000	64.0000/ 64.0000	0.2133/ 0.6400

its derivation relies on the concepts, axioms, and properties proposed by Esteban & Ray (1994). In this section, we will show that our measurement can be partly viewed as a (multi-dimensional) polarization measurement in the Wolfson's line.

2.6.1. INCREASED SPREAD AND INCREASED BIPOLARITY

The construction of a bipolarization measurement relies on two critical properties: **increased spread** and **increased bipolarity** (Wang & Tsui, 2000; Chakravarty & Majumder, 2001; Gigliarano & Mosler, 2009). Increased spread states that given the median level fixed, polarization increases when any individual moves in the opposite direction from the median level (Wang & Tsui, 2000), and increased bipolarity states that after a Pigou–Dalton transfer within the same group, polarization level should increase (Wang & Tsui, 2000; Gigliarano & Mosler, 2009). A Pigou-Dalton transfer is defined as a transfer from a rich individual to a poor individual, and after the transfer, the poor should not be richer than the rich before the transfer and the rich should not be poorer than the poor before the transfer (Wang & Tsui, 2000).

To see its relation with bipolarization measurements, we need to check if our ESBG-based measurement satisfies increased spread and increased bipolarity. From the definition of Pigou-Dalton transfer, it follows that W will be reduced after a transfer. However, estimating the effect of a Pigou-Dalton transfer on B without knowing the exact expression of B is not easy. For B defined in equation (2.13), a Pigou-Dalton transfer has no impact on it as the locations of the centroids are not affected. Therefore, at least the K-means-based polarization measurement proposed in Section 2.5.3 satisfies increased bipolarity.

Whether an ESBG-based measurement satisfies increased spread is a more complicated question. A data point's moving away from the median value (hereafter referred to as *increased-spread-move*) will definitely increase B. Meanwhile, depending on the location of the data point, the move may either increase or decrease W. Therefore, we do not know if P goes up or not. Some counter-examples can be found. For f^* given in Section 2.5.3, if we move the leftmost data point in X^* whose Variable 1 equals 0 to a more left location where Variable 1 is -10, by setting $\delta = 101$, the polarization measurement of the system drops from 0.003438371 to 0.003024571, mainly due to the significant increase in W. Although we are not sure if other expressions of $f(W,B)$ would satisfy increased spread, we could claim that this property is not generally desired by ESBG-based measurements.

2.6.2. IS THE ESBG-BASED MEASUREMENT A BIPOLARIZATION MEASUREMENT?

Even if our measurement may not satisfy increased spread, one cannot deny that it is similar to a bipolarization measurement in many aspects. First, ESBG itself is the same as the grouping method of a bipolarization measurement when $D = 1$ (see Section 2.4.1). This finding is interesting: we were looking for an appropriate grouping method for the Esteban & Ray's line, but after exploration, we end up with a grouping method similar to the one used in the Wolfson's line.

Secondly, the Wolfson's index (Wolfson, 1994) – the representative of the Wolfson's line – can also be written in the form of a function of W and B. The index is originally written in the following form (Wolfson, 1994; Wang & Tsui, 2000):

$$P^W = 2\frac{(2T - Gini)}{(m/\mu)} \tag{2.15}$$

where $T = 0.5 - L(0.5)$, and $L(0.5)$ is the share of variable of interest of the lower half of the population. $Gini$ refers to the Gini index of the whole population, m is the median value, and μ is the mean value (Wolfson, 1994). According to Gigliarano & Mosler (2009), the Wolfson's index can also be written as:

$$P^W = \frac{2\mu}{m}(B - W) \tag{2.16}$$

where W and B are represented by the Gini index between and within groups respectively (Gigliarano & Mosler, 2009). In this sense, both the ESBG-based measurement and the Wolfson's index are in the form of $P = f(W, B)$ (note that the Wolfson's index also depends on μ and m), and both of them are increasing with B and decreasing with W.

Finally, as shown in Section 2.6.1, an ESBG-based measurement – at least a particular form of it – satisfies increased bipolarity, one of the two basic properties of bipolarization measurements.

Since the ESBG-based measurement is not expected to satisfy increased spread, it should not be considered as a real bipolarization measurement. Given the similarities between them, we can roughly view the EBSG-based measurement as a (multi-dimensional) quasi-bipolarization measurement without the property of increased spread.

2.6.3. SQUEEZING-AND-MOVING FRAMEWORK

Polarization is a slippery, context-dependent concept whenever applied to social systems. Although we could understand in principle what a maximum or a minimum polarization is, the whole range of in-between states remains poorly understood. This explains why polarization is often described as the distance to the situation of maximum polarization. For instance, Flache and Mäs (2008) stated that "polarization captures the degree to which the group can be separated into a small set of factions who are mutually antagonistic in the opinion space and have maximal internal agreement". Indeed, the general interest in polarization, whether from the public or from scholars, mainly comes from the fear of its destructive effect on social harmony and stability (e.g., Layman & Carsey, 2002; Montalvo & Reynal-Querol, 2005; Fischer & Mattson, 2009). Such a fear-based interest naturally leads us to consider more carefully "how far are we from the most polarized situation?" rather than "what on earth is polarization?".

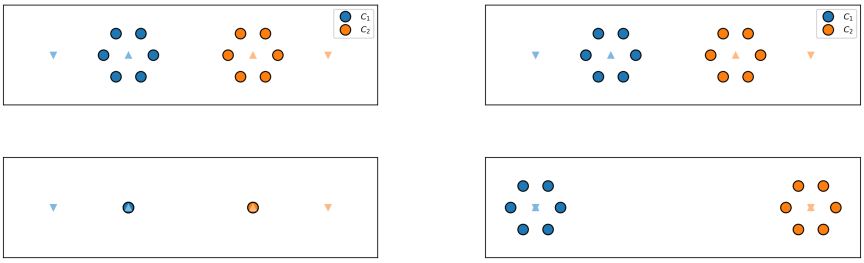
At a first glance, ESBG seems too simple to be correct. However, if we interpret polarization level as a measurement of "how far we are from the most polarized situation", it becomes clear why ESBG works. Consider that we want to transform a not-very-polarized data set into the maximum polarized situation. Therefore, the priority is to identify which data point should be relocated to which extreme. This is exactly what ESBG does.

To achieve maximum polarization, data points in each group should be later relocated to the nearer extreme. This task can be done via the following two steps: the **squeezing step** that "squeezes" the data points in the same group to the group center (Figure 2.12(a)), and the **moving step** that moves[10] each group to its corresponding extreme (Figure 2.12(b)). Given a group structure, within-group heterogeneity W measures how difficult the squeezing step is, and between-group heterogeneity B measures how easy the moving step is. This also explains why polarization should increase with B and decrease with W, if we consider polarization measurement as an index of the overall difficulty of achieving maximum polarization.

The concepts of squeezing and moving can help us to understand why the ESBG-based measurement and bipolarization measurement both satisfy increased bipolarity, but only the latter satisfies increased spread. A Pigou-Dalton transfer, by definition, will make the squeezing process easier (i.e., reducing W) without affecting the moving process (at least for f^* as the centroids are not affected by the transfer). Therefore, it facilitates the task and hence increases polarization. Therefore, both types of measurements satisfy increased spread.

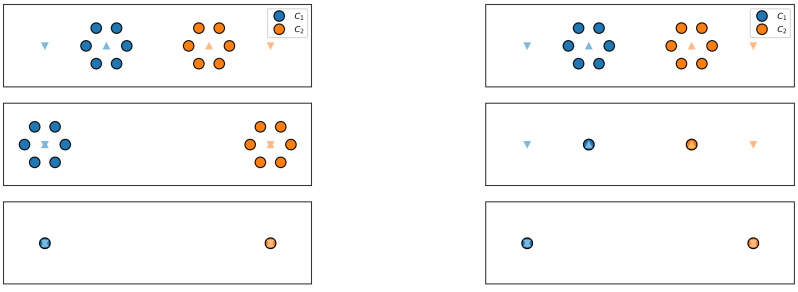
When considering increased spread, the picture is different. If the moving step is executed before the squeezing step (i.e., the moving-squeezing procedure, see Figure 2.13(a)), an increased-spread-move makes the moving process easier (i.e., increasing B) without affecting the squeezing process (see Figure 2.14(a)). It will therefore increase polarization. However, if the moving step is executed after the squeezing step (i.e., squeezing-moving procedure, see Figure 2.13(b)), an increased-spread-move makes the squeezing process more difficult (i.e., increasing W), while (maybe slightly) facilitating the moving process (i.e., increasing B) because the relevant centroid will be closer to its extreme due to the move (see Figure 2.14(b)). This implies that we cannot determine if the move will decrease polarization or not without knowing the exact expression of $f(W,B)$. From a result-oriented point of view, we can then conceptualize the bipolarization measurement as a realization of the moving-squeezing procedure, and the ESBG-based measurement as a realization of the squeezing-moving procedure, explaining why our measurement satisfies increased bipolarity but not increased spread.

[10] Strictly speaking, the "move" is a space translation that moves every point in the same group by the same distance and in the same direction.



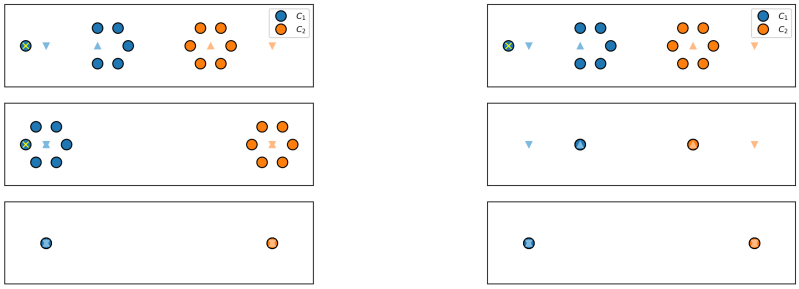
(a) Squeezing Step (b) Moving Step

Figure 2.12: Illustration of (a) the squeezing step and (b) the moving step. In each sub-figure, the configuration at the top will transfer to the configuration at the bottom after the step. The up-pointing triangles represent group centers (centroids) and the down-pointing triangles represent extremes.



(a) Moving-Squeezing Procedure (b) Squeezing-Moving Procedure

Figure 2.13: Illustration of (a) the moving-squeezing procedure and (b) the squeezing-moving procedure. In each sub-figure, the configuration at the top will transfer to the configuration at the bottom via the intermediate configuration in the middle. The up-pointing triangles represent group centers (centroids) and the down-pointing triangles represent extremes.



(a) Moving-Squeezing Procedure (b) Squeezing-Moving Procedure

Figure 2.14: Illustration of the increased-spread-move in (a) the moving-squeezing procedure and (b) the squeezing-moving procedure. The yellow saltire marks the data point that was moved here from the configuration at the top of Figure 2.13 (whether (a) or (b)) by an increased-spread-move. In each sub-figure, the configuration at the top will transfer to the configuration at the bottom via the intermediate configuration in the middle. The up-pointing triangles represent group centers (centroids) and the down-pointing triangles represent extremes.

2.7. CONCLUSION

In the vast literature on polarization, the notion of group, especially groups based on similarities between individuals, is the elephant in the room: everyone considers groups when defining or conceptualizing polarization, but it is difficult to understand what exactly such groups are. The only recurrent argument is that members of the same group should be similar, whereas members from different groups should be dissimilar. This is neither sufficient to capture the nuances of the various group structures, which are caused by various social cleavages that characterize our complex societies, nor it contributes to a consistent measurement of polarization. The mismatch between how we understand and how we measure polarization undermines the reliability of measurements, thus hampering our understanding of society in its complex and multifaceted aspects.

In this study, we have proposed a grouping method for constructing polarization measurements called "Equal Size Binary Grouping" (ESBG) that divides a data set into two groups of equal sizes according to similarities between data points. We showed that ESBG can be a suitable solution to certain theoretical and practical problems that trouble other grouping methods, such as discontinuity and contradiction of reasoning. While alternative approaches exist that over-impose pre-existing group structures or explore various dimensions of polarization, we believe that significant advances in polarization studies in complex societies can be made if measurements are consistent and possibly capable of discovering endogenous structures from data that are coherent with the variable(s) of interest.

Following clustering algorithms, we presented a procedure containing four steps to implement ESBG. Based on ESBG, a novel class of polarization measurements can be constructed to measure both uni- and multi-dimensional polarization. The measurements increase with between-group heterogeneity and decrease with within-group heterogeneity, and are not affected by other variables such as the number or size of groups. We also showed that the measurements satisfy a range of properties that have long been deemed desired in the field, such as continuity, normalization, maximization and minimization. Subsequently an illustrative example of applying ESBG and the related measurement to a synthetic data set was demonstrated.

As a final remark, we investigated the relation between the ESBG-based measurement and bipolarization measurement. The ESBG-based measurement can be roughly viewed as a multi-dimensional bipolarization measurement without the property of increased spread. This is because both types of measurements use the same grouping method when $D = 1$, and satisfy the same property of increased bipolarity. Furthermore, we developed a so-called "squeezing-and-moving" framework to help explain the relation between them.

With all due caveats due to our general approach and the lack of appropriate data on which to test these measurements, we believe that future research will help to improve the design of the measurement, while contributing to the debate on the key role of group definition in current measurements. Although useful to explore group structures within data starting from the variable(s) of interest, our method drastically simplifies the possible variety of groups co-existing in the same society, due to the varying cleavages that characterize the complex fabric of our social systems. However, we hope that our measurement could also stimulate new empirical research on polarization that improves comparability, replicability, and cumulativeness. As an avenue for future research, we suggest comparing

the ESBG-based measurement with existing polarization measurements in the context of various attribute distributions such as distributions with two or more peaks.

FUNDING

This study has received funding from the European Research Council: Consolidator Grant BEHAVE (Grant Agreement No. 724431).

REFERENCES

[1] Aaberge, R., & Brandolini, A. (2015). Multidimensional poverty and inequality. In A. B. Atkinson & F. Bourguignon (Eds.), Handbook of income distribution (Vol. 2, pp. 141-216). Elsevier.

[2] Abramowitz, A. I., & Saunders, K. L. (2008). Is polarization a myth?. The Journal of Politics, 70(2), 542-555.

[3] Aggarwal, C.C. (2014). An introduction to cluster analysis. In C. C. Aggarwal & C. K. Reddy (Eds.), Data clustering: Algorithms and applications (pp. 1–28). CRC Press.

[4] Aleskerov, F., & Oleynik, V. (2016). Multidimensional polarization index and its application to an analysis of the Russian state Duma. arXiv preprint arXiv:1608.01351.

[5] Alichi, A., Kantenga, M. K., & Sole, M. J. (2016). Income polarization in the United States. Working Paper, International Monetary Fund.

[6] Anderson, G. (2011). Polarization measurement and inference in many dimensions when subgroups can not be identified. Economics: The Open-Access, Open-Assessment E-Journal, 5.

[7] Ashman, K. M., Bird, C. M., & Zepf, S. E. (1994). Detecting bimodality in astronomical datasets. The Astronomical Journal, 108(6), 2348-2361.

[8] Baldassarri, D., & Bearman, P. (2007). Dynamics of political polarization. American Sociological Review, 72(5), 784-811.

[9] Baraldi, A., & Blonda, P. (1999a). A survey of fuzzy clustering algorithms for pattern recognition. I. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 29(6), 778-785.

[10] Baraldi, A., & Blonda, P. (1999b). A survey of fuzzy clustering algorithms for pattern recognition. II. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 29(6), 786-801.

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

[12] Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

[13] Boxell, L., Gentzkow, M., & Shapiro, J. M. (2017). Greater Internet use is not associated with faster growth in political polarization among US demographic groups. Proceedings of the National Academy of Sciences, 114(40), 10612-10617.

[14] Bramson, A., Grim, P., Singer, D. J., Fisher, S., Berger, W., Sack, G., & Flocken, C. (2016). Disambiguation of social polarization concepts and measures. The Journal of Mathematical Sociology, 40(2), 80-111.

[15] Bramson, A., Grim, P., Singer, D. J., Berger, W. J., Sack, G., Fisher, S., Sack, G., & Holman, B. (2017). Understanding polarization: Meanings, measures, and model evaluation. Philosophy of Science, 84(1), 115-159.

[16] Chakravarty, S. R., & Majumder, A. (2001). Inequality, polarisation and welfare: Theory and applications. Australian Economic Papers, 40(1), 1-13.

[17] Danzell, O. E., Yeh, Y. Y., & Pfannenstiel, M. (2019). Determinants of domestic terrorism: An examination of ethnic polarization and economic development. Terrorism and Political Violence, 31(3), 536-558.

[18] Deutsch, J., Fusco, A., & Silber, J. (2013). The BIP trilogy (bipolarization, inequality and polarization): One saga but three different stories. Economics: The Open-Access, Open-Assessment E-Journal, 7.

[19] DiMaggio, P., Evans, J., & Bryson, B. (1996). Have American's social attitudes become more polarized?. American Journal of Sociology, 102(3), 690-755.

[20] Duclos, J. Y., Esteban, J., & Ray, D. (2004). Polarization: concepts, measurement, estimation. Econometrica: Journal of the Econometric Society, 72(6), 1737-1772.

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

[22] Esteban, J., & Ray, D. (2012). Comparing polarization measures. In M. R. Garfinkel, & S. Skaperdas (Eds). The Oxford handbook of economics of peace and conflict (pp. 127–51). Oxford University Press.

[23] Esteban, J., & Schneider, G. (2008). Polarization and conflict: Theoretical and empirical issues. Journal of Peace Research, 45(2), 131-141.

[24] Duclos, J. Y., & Taptué, A. M. (2015). Polarization. In A. B. Atkinson & F. Bourguignon (Eds.), Handbook of income distribution (Vol. 2, pp. 301-358). Elsevier.

[25] Fiorina, M. P., & Abrams, S. J. (2008). Political polarization in the American public. Annual Review of Political Science, 11, 563-588.

[26] Fischer, C. S., & Mattson, G. (2009). Is America fragmenting?. Annual Review of Sociology, 35, 435-455.

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

[28] 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.

[29] 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).

[30] Forchheimer, D., Forchheimer, R., & Haviland, D. B. (2015). Improving image contrast and material discrimination with nonlinear response in bimodal atomic force microscopy. Nature Communications, 6(1), 1-5.

[31] Forgy, E. (1965). Cluster analysis of multivariate data: efficiency vs. interpretability of classifications. Biometrics, 21, 768-780.

[32] Foster, J. E., & Wolfson, M. C. (2010). Polarization and the decline of the middle class: Canada and the US. The Journal of Economic Inequality, 8(2), 247-273.

[33] Fusco, A., & Silber, J. (2014). On social polarization and ordinal variables: The case of self-assessed health. The European Journal of Health Economics, 15(8), 841-851.

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

[35] Hare, C., & Poole, K. T. (2014). The polarization of contemporary American politics. Polity, 46(3), 411-429.

[36] Hart, J. (1974). Symmetry and polarization in the European international system, 1870-1879: A methodological study. Journal of Peace Research, 11(3), 229-244.

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

[38] Homan, A. C., Van Knippenberg, D., Van Kleef, G. A., & De Dreu, C. K. (2007). Interacting dimensions of diversity: Cross-categorization and the functioning of diverse work groups. Group Dynamics: Theory, Research, and Practice, 11(2), 79-94.

[39] Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data. Prentice-Hall.

[40] Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. ACM Computing Surveys (CSUR), 31(3), 264-323.

[41] Kim, J., & Baek, S. K. (2021). Democracy and polarization in the National Assembly of the Republic of Korea. Journal of the Korean Physical Society. arXiv preprint arXiv:2101.03490.

[42] Knapp, T. R. (2007). Bimodality revisited. Journal of Modern Applied Statistical Methods, 6(1), 8-20.

[43] Layman, G. C., & Carsey, T. M. (2002). Party polarization and "conflict extension" in the American electorate. American Journal of Political Science, 46(4), 786-802.

[44] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (pp. 281-297).

[45] 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.

[46] McCright, A. M., & Dunlap, R. E. (2011). The politicization of climate change and polarization in the American public's views of global warming, 2001–2010. The Sociological Quarterly, 52(2), 155-194.

[47] Montalvo, J. G., & Reynal-Querol, M. (2005). Ethnic polarization, potential conflict, and civil wars. American Economic Review, 95(3), 796-816.

[48] Müller, A. C., & Guido, S. (2016). Introduction to machine learning with Python: A guide for data scientists. O'Reilly Media.

[49] Phillips, K. W. (2003). The effects of categorically based expectations on minority influence: The importance of congruence. Personality and Social Psychology Bulletin, 29(1), 3-13.

[50] Ross, E. A. (1920). The principles of sociology. Century Company.

[51] Schweighofer, S., Schweitzer, F., & Garcia, D. (2020). A weighted balance model of opinion hyperpolarization. Journal of Artificial Societies and Social Simulation, 23(3).

[52] Takács, K., Bravo, G., & Squazzoni, F. (2018). Referrals and information flow in networks increase discrimination: A laboratory experiment. Social Networks, 54, 254-265.

[53] Wang, Y. Q., & Tsui, K. Y. (2000). Polarization orderings and new classes of polarization indices. Journal of Public Economic Theory, 2(3), 349-363.

[54] Wolfson, M. C. (1994). When inequalities diverge. The American Economic Review, 84(2), 353-358.

[55] Xu, R., & Wunsch, D. C. (2009). Clustering. Wiley. https://doi.org/10.1002/9780470382776

[56] Zhang, X., & Kanbur, R. (2001). What difference do polarisation measures make? An application to China. Journal of Development Studies, 37(3), 85-98.

[57] Zhang, C., Mapes, B. E., & Soden, B. J. (2003). Bimodality in tropical water vapour. Quarterly Journal of the Royal Meteorological Society, 129(594), 2847-2866.

3

Learning Opinions by Observing Actions:Simulation of opinion dynamics using an Action-Opinion Inference model

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).**
DOI:10.18564/jasss.4020

Opinion dynamics models are based on the implicit assumption that people can observe the opinions of others directly, and update their own opinions based on the observation. This assumption significantly reduces the complexity of the process of learning opinions, but seems to be rather unrealistic. Instead, we argue that the opinion itself is unobservable, and that people attempt to infer the opinions of others by observing and interpreting their actions. Building on the notion of Bayesian learning, we introduce an Action-Opinion Inference model (AOI model); this model describes and predicts opinion dynamics where actions are governed by underlying opinions, and each agent changes her opinion according to her inference of others' opinions from their actions. We study different action-opinion relations in the framework of the AOI model, and show how opinion dynamics are determined by the relations between opinions and actions. We also show that the well-known voter model can be formulated as being a special case of the AOI model when adopting a bijective action-opinion relation. Furthermore, we show that a so-called inclusive opinion, which is congruent with more than one action (in contrast with an exclusive opinion which is only congruent with one action), plays a special role in the dynamic process of opinion

spreading. Specifically, the system containing an inclusive opinion always ends up with a full consensus of an exclusive opinion that is incompatible with the inclusive opinion, or with a mixed state of other opinions, including the inclusive opinion itself. A mathematical solution is given for some simple action-opinion relations to help better understand and interpret the simulation results. Finally, the AOI model is compared with the constrained voter model and the language competition model; several avenues for further research are discussed at the end of the paper.

Keywords: Opinion dynamics, Norm formation, Voter model, Behavioral change

3.1. INTRODUCTION

The study of opinion dynamics is a well-established topic in socio-physics (Castellano et al., 2009), which has continued to attract the attention of scholars for more than two decades. A vibrant research community of physicists and social scientists has shown increasing interest in describing opinion spreading, as a potential determinant of norm formation and behavioral change, by modeling the interactions of virtual agents in social networks. Various models and corresponding simulations have been proposed to explain how consensus or diversity is reached for a group of interacting agents with different opinions (Vazquez et al., 2003; Svenkeson & Swami, 2015). Among these models, the voter model serves as one of the simplest, most elegant, and most well-known examples of completely-solved opinion dynamics (Sood & Redner, 2005). It assumes that an agent would adopt "the opinion of a randomly chosen neighbor" (Krapivsky & Redner, 2003). Based on this minimal assumption, the voter model describes the formation of global consensus (Krapivsky et al., 2010) and also produces fruitful non-trivial behaviors when extended or modified (Mobilia, 2003; Vazquez et al., 2003; Lambiotte & Redner, 2007; Wang et al., 2014), or applied to various network forms (Castellano et al., 2003; Suchecki et al., 2005). Besides the voter model, other notable models of opinion dynamics, just to name a few, includes the Sznajd model (Sznajd-Weron, 2005), social influence model (Flache et al., 2017) and contagion model (Valente, 1996; Pacheco, 2012). Each of them has captured important aspects of opinion dynamics.

Regardless of the differences among these models, they predominantly assume that agents have the ability to directly observe the opinions of other agents, and then update their own opinions according to the observation. This notion of "*learning opinions by observing opinions*" is based on the implicit assumption that opinions are directly observable; or alternatively, that, when agents interact, they ask for each other's opinion, and express it to each other – truthfully – when asked for. However, in most situations, and particularly so for sensitive topics, such assumptions seem to be unrealistic: what we can learn from everyday life is that opinions are usually unobservable, and talking about opinions is not the most common way for people to exchange them. Although the term "opinion" may be used in a generic way to represent the property (e.g., attitude, belief, evaluation) affected by others (Flache et al., 2017), there is no question about the latency of opinions. Instead of learning opinions directly, we usually infer other people's opinions by interpreting their actions. Actions can be, for instance, choosing between cycling and driving for one's daily commute. These actions are observable, but the opinions that induce them are not. However, the relation between opinions and actions helps us learn

the unobservable opinions by interpreting observable actions. For example, an observer might interpret someone's choice to cycle to work as evidence of an environmentally-conscious mindset. The action-opinion relation, however, is not always clear: sometimes the observer is confused about which opinion in fact governs the observed action. For instance, the cyclist could simply be highly cost-sensitive, or care about health a lot, while not caring about the environment at all. Likewise, believers in one opinion can take various actions: both cycling and driving an electric car are reasonable options for an environmentalist (Moons & De Pelsmacker, 2012). The possible multiplicity in the relations between opinions and actions has not been captured by previous models yet.

In social psychology, the relation between opinions and actions has been extensively studied with a different term called *attitude-behavior consistency/ inconsistency*, where *attitudes* resemble *opinions*, and *behaviors* resemble *actions* in this paper. Intuitively, early studies assumed that "attitudes predict overt behaviors" (Zanna et al., 1980), but this notion has been challenged by psychologists who found that the attitude-behavior relationship is considerably weak (Wicker, 1969), and may be influenced by other factors (Liska, 1984). It is called the problem of *attitude-behavior inconsistency*. One of the most influential and successful models that aims to explain the problem should be the Fishbein-Ajzen model (Fishbein & Ajzen, 1975). The model proposes that behavior is directly driven by behavioral intention, and intention depends on both attitudes and subjective norms (Liska, 1984), providing a conceptualized framework to analyze multiplicity in attitude-behavior relations, which might shed light on opinion-action relations that will be used in our paper.

In this paper, we propose a simple alternative model of opinion-learning to simulate opinion dynamics in an artificial society. The central assumption embedded in this so-called Action-Opinion Inference (AOI) model is the notion of "*learning opinions by observing actions*", which means that the agent attempts to learn the opinions of others by observing and interpreting their actions. Crucially, our model assumes that actions are noisy signals of underlying opinions, which follows from the above-discussed multiplicity. Agents try to infer opinions from actions in a Bayesian way, acknowledging the uncertainty inherent to the opinion-action relation. More specifically, acknowledging the multiplicity of action-opinion relations, our model is able to describe the situation where an action is permitted by more than one opinion, and one opinion could result in various actions. Therefore observing an action will in most cases only allow the observer to partially update her beliefs about the other agent's underlying (latent) opinions. The model postulates that agents update their opinions and actions by a three-stage mechanism: the agent first uses Bayes' rule to update her beliefs regarding her neighbors' opinions, based on their actions which she observes. The inference of opinions from observed actions is determined by an action-opinion matrix, which defines, for each action and each opinion, if the action is either prohibited, permitted, or obliged by that particular opinion. In the second stage, the agent chooses her new opinion for the next time step according to the relative probability of each opinion in the neighborhood, calculated from the inferences of different opinions. In stage three she updates her action according to the new opinion she selected just now. Having performed an extensive literature review, we only found one model whose conceptualization of latent opinions and observable actions relates to our work in physics. This so-called CODA model (Martins, 2008) and its relation to our

work will be described at the end of Section 3.3.

We compare the AOI model systematically with the voter model, which we will prove can be formulated as a reduced form of the AOI model. More specifically, the AOI model is equivalent to the voter model when each action is only obliged by one opinion, and each opinion only obliges one action (when there is no uncertainty as to which opinion causes which action). The model setup offers an opportunity for us to explore different action-opinion relations, which have rarely been studied before. In the paper we employ cellular automata (CA) to simulate the model with different action-opinion relations and analyze the simulation results both numerically and spatially, focusing on the density of each opinion in the final stable state, as well as the clustering features of the dynamic process. The key question we investigate in this paper is as follows: "what action-opinion relations induce consensus or diversity?" In other words, we would envisage the role of the action-opinion relations in the formation of macroscopic features of the society. This question is closely related to the well-known Axelrod's question, which asks why consensus is not always reached given that agents learn opinions from others (Axelrod, 1997). In this sense, the AOI model provides an alternative approach to answering Axelrod's question besides the conventional models mentioned before.

The rest of the paper is organized as follows: **Section 3.2 Public & Private Characteristics: A Brief Review**, as the name indicated, briefly reviews the relevant works on public and private characteristics, a similar concept to our notion of *learn opinions by observing actions*. **Section 3.3 Model Setup** presents the Action-Opinion Inference model in detail. **Section 3.4 Two-Action Situation** and **Section 3.5 Three-Action Situation** illustrate the simulation results of the model with two and three actions respectively. Additionally, a brief mathematical analysis is given in **Section 3.4** for the system with two opinions and two actions only. The mathematical analysis provides the first step towards validation of the simulation results, and helps us better understand the dynamic process. **Section 3.6 Conclusion and Discussion** summarizes the major findings and discusses some critical issues concerning the AOI model and other complex system models. Furthermore, we discuss several avenues for further research.

3.2. PUBLIC & PRIVATE CHARACTERISTICS: A BRIEF REVIEW

Although the notion of *learning opinions by observing actions* was not frequently acknowledged in previous literature, a similar pair of concepts – private and public characteristics – has been employed in previous opinion dynamics studies to capture the discrepancy in the learning process. Here, the term "characteristics" may refer to opinions, attitudes, actions, or any property of an agent that is open to the influences from others. "Public characteristic" represents the observable characteristic publicly expressed by an agent. Conversely, a private characteristic is defined as an agent's privately held characteristic. Therefore, a public characteristic can be observed directly but is not necessarily the same as the agent's private characteristic. An early example comes from **information cascade**. Information cascade, which is defined as the situation when agents simply follow the actions of the others sequentially without considering their own private information (Bikhchandani et al., 1992), is a powerful tool to explain localized conformity and its systematical fragility (Bikhchandani et al., 1992; Wu & Huberman, 2004). Both information cascade and the AOI model have roots in the same idea that the public characteristic

obtained from a neighbor may not be the same as the neighbor's private characteristic, and the agent takes actions by making inference from that possibly inaccurate public characteristic. In our notation, an action serves as the public characteristic, and the opinion beneath is the true but private characteristic. The discrepancy between public and private characteristic in information cascade originates from the fact that agents simply ignore their own private characteristic (i.e., opinions) when taking actions, but in the AOI model, it is because of the unobservable relations between actions (public characteristic) and opinions (private characteristic). Therefore the two notions have different underlying principles.

A more recent representative of public and private characteristics is the **persuasion model** (Mäs et al., 2013; Mäs & Flache, 2013; Mäs & Bischofberger, 2015) based on psychological theories (e.g., Fishbein, 1963; Petty et al., 1981). Different from traditional opinion dynamics models, the persuasion model assumes that opinions are formed based on arguments, and agents only exchange arguments, so opinions are not directly influenced by others (Mäs & Bischofberger, 2015). One may realize that the underpinning of our AOI model is inherently close to this assumption in the persuasion model. In both models, opinions (i.e., the private characteristic) play no role in the communication (not necessarily verbal) directly, while some other public characteristics, which refer to actions in the AOI model and arguments in the persuasion model, serve as the messenger between agents. In the persuasion model agents learn arguments from others, and form new opinions based on the arguments; Meanwhile, in the AOI model agents observe actions of others, and update their opinions according to the interpretation of the observations. The primal difference between the two models lies in the relations between private characteristic (i.e., opinions) and public characteristic (i.e., action or argument): in the persuasion model, opinion is a function of some relevant arguments, and thus arguments can affect opinions, but not vice versa. In other words, an agent's opinions are only affected by another agent's arguments. On the contrary, the opinion in the AOI model, together with the action-opinion matrix, determines the action; and only another agent's actions, via the inference process, can affect the agent's opinion. That is, *actions* are a function of opinions. This structural disparity distinguishes the two models at a microscopic level, and thus will lead to distinct outcomes at a macroscopic level (see further below). It is noteworthy that in practice, persuasion models usually adopt *opinion homophily*, that is, each agent selects an agent she wants to interact with based on the similarities of their opinions (Mäs & Flache, 2013), so opinions are still observable in such models, playing the role of labels in the partner selection phase. Opinion homophily should be partly responsible for the persuasion model's ability to explain opinion polarization.

Another famous phenomenon describing disparities between public and private characteristics is **pluralistic ignorance**, in which most members of a society privately disapprove of, or are undecided about, an opinion but incorrectly believe that most other members accept it (Miller & McFarland, 1987; Huang & Wen, 2014). Considering themselves as the only dissident, they would express their approval of the opinion that they do not actually support. Pluralistic ignorance results in a global consensus although most members disagree with it, and hence the consensus is so fragile that it could be broken by the so-called *minority influence* (Huang & Wen, 2014). As Seeme and Green explained, the term "opinion" in studies of pluralistic ignorance, rather than in the AOI

model, refers to *"the expression or behavior of a person towards a topic"*, instead of one's *"true internal opinion"* (Seeme & Green, 2016). The "true internal opinion" is called "attitude" by Seeme & Green (2016) to avoid confusion. According to this claim, we find that the AOI model and the pluralistic ignorance studies both describe the discrepancy of the observable "public characteristic", which is *actions* in the AOI model and publicly expressed *opinions* in pluralistic ignorance, and the "private characteristic", which refers to the (underlying) opinions in the AOI model and (private) attitudes in pluralistic ignorance. The critical difference between these studies is obvious: the dynamics of public characteristic (*actions*) in the AOI model is driven by the dynamics of private characteristic (*opinions*). As stated in Section 3.2, public characteristics (*actions*) are not directly influenced. However, in pluralistic ignorance models, public characteristics are directly affected by "the pressure to conform" (Seeme & Green, 2016) or "normative social influence" (Huang & Wen, 2014), and private characteristics (*attitudes*) are later updated according to either "self-perception theory", "cognitive dissonance" (Seeme & Green, 2016), or other psychological theories.

Table 3.1 summarizes the three models involving discrepancies between public and private characteristics as well as the AOI model itself. One of the relevant references for each model is listed inside the parentheses below.

Table 3.1: Comparison of models with public characteristics (PC) & private characteristics (PrC)

Model	PC	PrC	Dynamics of PC	Dynamics of PrC
AOI (this paper)	Action	Opinion	Driven by PrC	Observe-Infer
Information Cascade (Bikhchandani et al., 1992)	Action	Opinion	Observe-Infer	Ignored by agents
Persuasion Model (Mäs & Bischofberger, 2015)	Argument	Opinion	Exchange	Driven by PC
Pluralistic Ignorance (Seeme & Green, 2016)	Opinion	Attitude	Pressure to conform	Psychological theories

The table shows the similarities as well as differences between some notable earlier work and the AOI model. In all, although previous researches have noticed the existence of public and private characteristics and described them in various models, these models do not capture the inference process that enables agents to learn the private characteristics of others by observing public characteristics. The absence of an inference process leads to the omission of uncertainty: in these three models, the relations between public and private characteristics are either deterministic (persuasion model & pluralistic ignorance) or unspecified (information cascade). In contrast, the AOI model creates a smokescreen between public characteristics (*actions*) and private characteristics (*opinions*), which represents the multiplicity of the action-opinion relations. This type of uncertainties, although rarely acknowledged in opinion dynamics papers, could lead to misunderstanding or obfuscation, and the role of the uncertainties in opinion dynamics will be one of the central problems we investigate in the rest of the paper.

3.3. MODEL SETUP

We consider a population of N agents on an $L \times L$ regular lattice with periodic boundary conditions as well as a Von Neumann neighborhood. Each cell of the lattice is occupied by an agent, and we set $L^2 = N$ to avoid empty cells. Agent i ($i = 1, 2, ..., N$) chooses one action $a^{(i)}$ from the action set $A = \{a_1, ..., a_g, ..., a_G\}$ based on her opinion, described by a rule $r^{(i)}$ selected from the rule set $R = \{r_1, ..., r_k, ..., r_K\}$. Note that, in terms of terminology, we choose to use the term "opinion" in colloquial discussions, and we use the term "rule" in the context of the mathematical model and simulations. The evaluation of action a_g by rule r_k is denoted by s_{kg}. In case r_k is an **exclusive rule**, $s_{kg} \in \{+, -\}$ a_g is either obliged ($+$) or prohibited ($-$) by r_k; however, if r_k is an **inclusive rule**, then $s_{kg} \in \{0, -\}$, where 0 implies that the action is permitted but not obliged by the rule. An exclusive rule can only oblige one action, but an inclusive rule always permits more than one action. All s_{kg} ($g = 1, ..., G$; $k = 1, ..., K$) constitute a $K \times G$ matrix S, called the action-opinion matrix, summarizing the action-opinion relations in the system.

The behaviors of the agents are described as follows: if an agent follows an exclusive rule r_k, then she will certainly take the action obliged by the rule, that is, $P(a_g|r_k) = 1$ if $s_{kg} = +$. Otherwise $P(a_g|r_k) = 0$. If she believes in an inclusive rule r_k, then $P(a_g|r_k) = 1/W$ if $s_{kg} = 0$, where W is the number of actions permitted by r_k; $P(a_g|r_k) = 0$ if a_g is prohibited (i.e., $s_{kg} = -$). Besides, the agent can observe the actions of her Von Neumann neighbors but cannot observe their opinions (in the form of rules) directly. In addition, each agent has full knowledge of A, R, S, and she assumes that other agents choose actions and update rules in the same way as she does so herself.

Proceeding on the preliminary setup, the action-opinion inference process takes the following steps. Initially, each agent (say agent i) is randomly assigned a rule $r^{(i)} \in R$, then chooses the action based on the assigned rule. At each time unit τ, an agent (say agent i) is randomly chosen to update her probabilistic inference about which rule is adopted by her neighbor j ($j \in M_i$, M_i is the von Neumann neighbors of agent i) based on the observation of j's action $a^{(j)}(\tau)$. Specifically, agent i's inference that neighbor j adopts r_k after observing $a^{(j)}(\tau)$ takes the form:

$$P^{(i)}(r^{(j)}(\tau) = r_k | a^{(j)}(\tau)) = \frac{P(a^{(j)}(\tau)|r_k)}{\sum_{k=1}^{K}[P(a^{(j)}(\tau)|r_k)]} \tag{3.1}$$

where $P(a^{(j)}(\tau)|r_k)$, which has been defined above, is the probability that an agent acts as $a^{(j)}$ given the rule r_k. It should be noted that all the agents have the same inference strategy, thus $P^{(i)}(r^{(j)}(\tau) = r_k|a^{(j)}(\tau)) = P^{(l)}(r^{(j)}(\tau) = r_k|a^{(j)}(\tau))$ $\forall \tau, k$ if $i, l \in M_j$.

Equation (3.1) is derived from the Bayes' rule by setting equal prior probabilistic beliefs $P(r_k)$ for all k, that is, $P(r_k) = 1/K$, $\forall k$. The intuition behind this is that the agent *a priori* assumes that each rule is equally likely to be taken by her neighbor j. This assumption is reasonable in light of the fact that in each time unit only one agent is selected to observe the neighborhood and then update rules and actions, so it is highly likely that she did not observe her neighbor's action in a recent time unit.

By updating her probabilistic inference of all neighbors' rules, agent i learns the local opinion distributions, based on which she will update her own opinion (i.e., rule). To do

so, agent i evaluates the accumulative probability of each rule across the neighborhood:

$$\hat{P}^{(i)}(r_k,\tau) = \sum_{j\in M_i} P^{(i)}(r^{(j)}(\tau) = r_k | a^{(j)}(\tau)) \tag{3.2}$$

for $k = 1,2,...,K$. The normalized probability set $\{\hat{P}^{(i)}(r_k,\tau)/\sum_k \hat{P}^{(i)}(r_k,\tau)\}_{k=1,2,...,K}$ helps agent i to estimate the occurrence of each rule. Agent i will adopt r_k as her rule for the next time unit with the probability $\hat{P}^{(i)}(r_k,\tau)/\sum_k \hat{P}^{(i)}(r_k,\tau)$.

After the rule-updating process, the world moves to the next time unit $\tau+1$. Because in each time unit only one agent updates, it is inefficient to study the dynamics between time units. Instead, we denote N consecutive time units as one time step t so that each agent has been selected once during one time step on average. This is a common practice adopted by many opinion dynamics models (Suchecki et al., 2005). We will present the dynamics of key variables in the scale of time step in the rest of the paper.

The voter model is chosen as the benchmark for the AOI model not only because of its long-standing popularity in the discipline of opinion dynamics for explaining the emergence of consensus (Sood & Redner, 2005; Barrat et al., 2008; Krapivsky et al., 2010), but also due to the fact that the AOI model is built upon the framework of the voter model: despite the AOI model's learning process, the basic dynamics of the two models are the same, thus using the voter model as a benchmark helps derive implications regarding the effect of the inference. Moreover, variations of the voter model have been applied to a wide range of social phenomena besides opinion dynamics, which provides examples for the AOI model to be modified for other disciplines. The reason why we use the voter model as the basis of the AOI model is also related to its simplicity. As argued by many, the voter model is the simplest and minimal model for the study of opinion dynamics, so its basic framework helps maintain the simplicity and comprehensibility of the new model. For example, if the social influence model were chosen as the basis, we might have trouble deciding which type of social influence we would like to use – positive, negative, assimilative, or similarity biased. Combining the action-opinion inference with various social influence models is a promising line of further research, but for the first work on the AOI model, it is better to avoid unnecessary subtleties arising from model setup and focus on the role of action-opinion inference process in governing dynamics.

The voter model is an example of discrete opinion models, where opinion is represented by a discrete variable. Therefore in the AOI model, rules and actions are also discrete. Besides discrete opinion models (e.g., voter model, majority rule model (Galam, 2002), Sznajd model (Sznajd-Weron, 2005)), many sociological models describe variables of interest in a discrete way, including Latané's social impact theory (Latané, 1981) and Axelrod's model of cultural dissemination (Axelrod, 1997). Continuous opinion models have taken an alternative approach, where opinions can vary between extreme values smoothly (Castellano et al., 2009). Deffuant model (Deffuant et al., 2000), Hegselmann-Krause model (Hegselmann & Krause, 2002) and social influence models (Flache et al., 2017) are famous examples of continuous opinion models. A model, mentioned in the introduction, which shares with our model its distinction between latent opinions and observable actions, is the so-called CODA model (Martins, 2008). That model differs from ours in the following fundamental ways: the CODA model postulates that opinions refer to a ground (or: universal) truth, which agents attempt to uncover by learning from each

others' actions in a Bayesian fashion. In contrast, the AOI model is not concerned with learning truths, but inferring the latent beliefs of neighbors, induced by a wish to conform with those neighbors in terms of their latent beliefs. This is why the AOI model postulates that opinions and actions are discrete and multinomial, as opposed to the CODA model whose actions are discrete and binary, and whose opinions are statements in the form of a continuous probability function about something being true or not. For this same reason, the AOI model distinguishes between an inference step in which agents try to infer (learn) each others' opinions, and an update step in which agents update (adjust) their own opinion as a probabilistic function of neighbors' opinions. This distinction is absent in the CODA model, which is logical given that it is concerned only with learning about truths. Conceptually, the CODA model is related to literature about truth-seeking agents (Prelec, 2004), whereas the AOI model is focused on opinion-conformity among agents with diverse latent opinions. An important distinction between the models in terms of what macro-level phenomena they tend to predict, is that the CODA model tends to generate extremism, whereas the AOI model, depending on the structure imposed on action-opinion relations may generate either consensus, coexistence of various opinions, or extremism. This difference in predicted outcomes is rooted in the fact that the CODA model features one relation between binary actions and an underlying opinion, whereas the AOI model features a broader set of relations between various actions and various opinions which may either be exclusive or inclusive (see the first paragraph of Section 3.3).

3.4. TWO-ACTION SITUATION

For simplicity, we first focus on the two-action AOI model (i.e., $G = 2$). Given a two-action set $\{a_1, a_2\}$, there are in total 4 possible action-opinion matrices S_1 to S_4 for $K \leq 3$, if one does not allow for duplication.

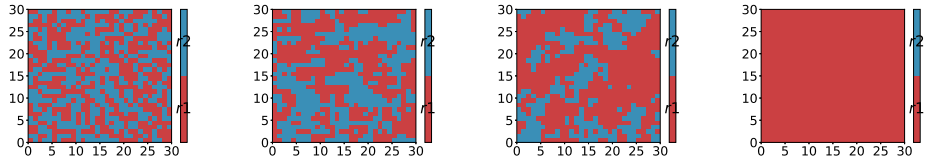
$$S_1 = \begin{array}{c} \\ r_1 \\ r_2 \end{array} \begin{array}{c} \begin{array}{cc} a_1 & a_2 \end{array} \\ \begin{pmatrix} + & - \\ - & + \end{pmatrix} \end{array} \qquad S_2 = \begin{array}{c} \\ r_1 \\ r_2 \end{array} \begin{array}{c} \begin{array}{cc} a_1 & a_2 \end{array} \\ \begin{pmatrix} + & - \\ 0 & 0 \end{pmatrix} \end{array} \qquad S_3 = \begin{array}{c} \\ r_1 \\ r_2 \end{array} \begin{array}{c} \begin{array}{cc} a_1 & a_2 \end{array} \\ \begin{pmatrix} - & + \\ 0 & 0 \end{pmatrix} \end{array} \qquad S_4 = \begin{array}{c} \\ r_1 \\ r_2 \\ r_3 \end{array} \begin{array}{c} \begin{array}{cc} a_1 & a_2 \end{array} \\ \begin{pmatrix} + & - \\ - & + \\ 0 & 0 \end{pmatrix} \end{array}$$

Note that S_2 and S_3 are identical in nature. We will study S_1, S_2 and S_4 only, which cover all action-opinion relations in a two-action situation when there are more than one rule.

3.4.1. AOI MODEL WITH S_1 AND THE VOTER MODEL

When taking S_1 as the action-opinion matrix for the system, the AOI model reduces to a two-state voter model, a naive spin model where agents observe and learn opinions directly. This serves as a simple but representative example of studies in opinion dynamics and consensus formation (Krapivsky et al., 2010). In the voter model, a randomly chosen agent adopts the opinion of a neighbor who is also chosen at random (Dornic et al., 2001). The voter model can therefore be interpreted as a special case of the AOI model, the two models being equivalent when in the AOI model, each action is only obliged by one rule, and each rule only obliges one action (i.e., S_1). Unsurprising, the AOI model with S_1 shows all the features of a two-state voter model (Figure 3.1). Starting with a random initial configuration, the population eventually converges to an absorbing state of consensus

where everyone takes the same opinion (and action). Also, the emergence of opinion clustering and coarsening shown in Figure 3.1 is a typical pattern of the voter model (Krapivsky et al., 2010).

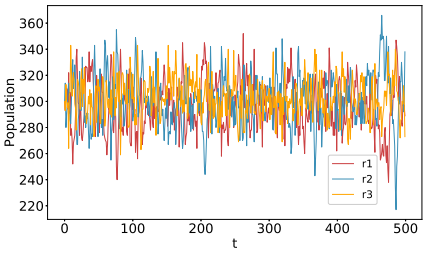


(a) $t = 1$ (b) $t = 5$ (c) $t = 50$ (d) $t = 500$

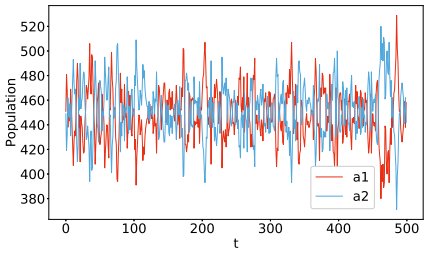
Figure 3.1: Snapshots of rule distribution of the AOI model with $S = S_1$ on an $L = 30$ lattice with a random initial configuration and equal densities of both rules. $A = \{a_1, a_2\}$, $R = \{r_1, r_2\}$.

3.4.2. SIMULATIONS OF AOI MODELS WITH S_4 AND S_2

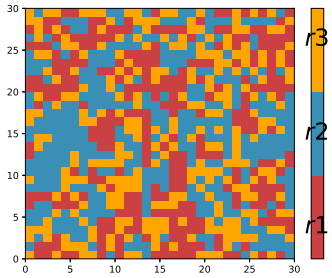
An inclusive rule (r_2 in S_2 and r_3 in S_4) that permits both actions is introduced in S_2 and S_4. S_4 constitutes a typical case in politics where both supporters of a party (believe in r_1 or r_2) and indifferent voters (believe in r_3) exist. See Section 3.6.1 and Section 3.6.2 for a discussion of how these indifferent voters can be seen as centrists. A striking observation from the simulation of the system with S_4 (Figure 3.2) is that the two key features of the voter model (i.e., the AOI model with S_1), clustering and consensus, are no longer valid when an inclusive rule is introduced (i.e., using S_4). Figure 3.3 shows the interface densities of rules or actions for both S_1 (voter model) and S_4. *Interface density*, sometimes called *density of domain walls*, is defined by the fraction of neighboring agents with different rules or actions (Krapivsky et al., 2010). Therefore a lower interface density indicates a higher level of clustering. We find that the inclusive rule r_3 significantly reduces the clustering of opinions (or actions) compared to the voter model. Meanwhile, the population reaches a dynamical disordered state, where all three rules coexist and the density of each rule remains relatively stable over time (Figure 3.2 (a) and (c)). Please be aware that the y-axis scales of (a) and (b) in Figure 3.2 are different, and other figures in the rest of the paper may show different sections of the scale.



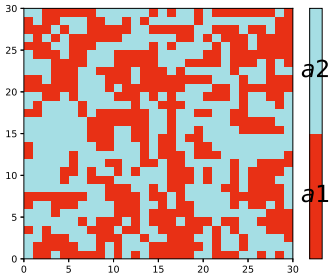
(a) Population dynamics of rules



(b) Population dynamics of actions



(c) Population distribution of rules at $t = 500$



(d) Population distribution of actions at $t = 1000$

Figure 3.2: Simulation results of the AOI model with $S = S_4$ on an $L = 30$ lattice with a random initial configuration and equal initial density of each rule. $A = \{a_1, a_2\}$, $R = \{r_1, r_2, r_3\}$.

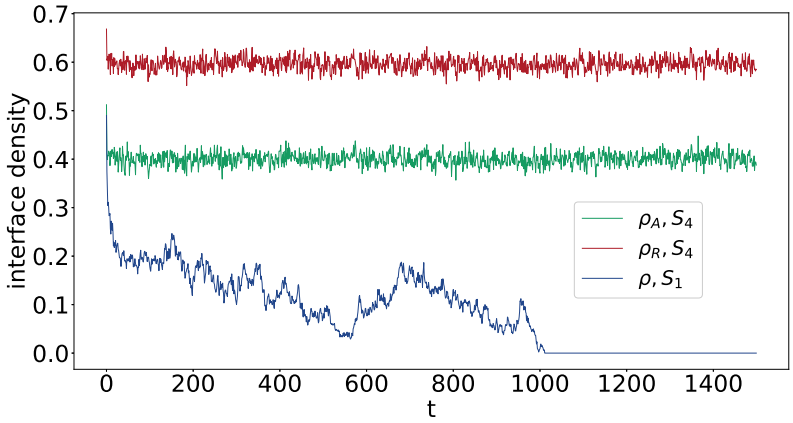


Figure 3.3: Evolution of the interface density for the AOI model with S_1 and S_4. For S_1, ρ is the interface density of rules (or equivalently, actions). For S_4, ρ_A is the interface density of actions, and ρ_R is the interface density of rules. Model setup: $L = 30$, random initial configuration and equal density for each rule.

Holley and Liggett (1975) have proved that coexistence of opinions is impossible in a two-dimensional voter model, and many efforts have been made to modify the model in order to reach a mixed state where more than one opinion survives. Such modifications include the threshold voter models (Liggett, 1994) where the agent adopts the opposite opinion only when the number of neighbors with opposite opinions is large enough, and the three-state constrained voter model (Vazquez et al., 2003), where the leftists and rightists only interact with the centrists. Additionally, if the voter model is run on a small-world network (Watts & Strogatz, 1998), the system will be temporarily trapped in a metastable state where different opinions coexist, although it will escape from the metastable state and reaches consensus eventually (Castellano et al., 2003; Castellano et al., 2009). In Figure 3.2 we have already found that the AOI model, which is an extension of the voter model, can reach the mixed state of opinions by simply introducing an inclusive rule (e.g., r_3 in S_4) without restricting the interactions of agents (as in the threshold and constrained voter models) or modifying the network structure (as in the small-world network).

The intuition behind the coexistence of opinions in Figure 3.2, which concerns S_4, is straightforward. For example, if an observer (the focal agent) sees a neighbor acting as a_1, she considers the neighbor believes in r_1 with probability $P(r_1|a_1) = \frac{1}{1+0.5} = \frac{2}{3}$, and r_3 with a smaller probability $P(r_3|a_1) = \frac{0.5}{1+0.5} = \frac{1}{3}$ according to S_4. So a neighbor acting as a_1 will increase the observer's probability of adopting r_1 as well as (albeit less so) r_3. Similarly, an action a_2 of a neighbor will not only increase the observer's probability of adopting r_2, but also increase her probability of adopting r_3. The underlying opinion dynamics imply that r_3 will never die out. Likewise, an agent employing r_3 will take a_1 and a_2 with equal probabilities, and therefore a reciprocal loop of opinions is constituted (Figure 3.4 (c)). The loop shows that each action or rule can reach any other action or rule through a finite number of arrows, which implies that all rules and actions are "beneficial" to all the others. This explains the coexistence of different opinions (and actions) in the AOI model with S_4. Contrarily, Figure 3.4 (a) shows that in S_1, r_1 and r_2 are disconnected, so the ultimate consensus is always reached.

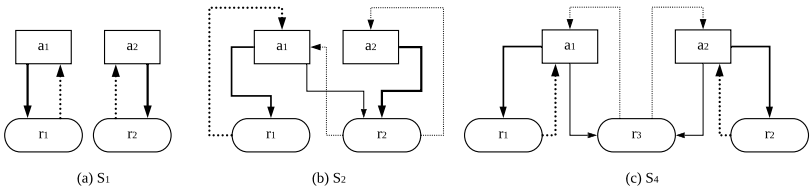
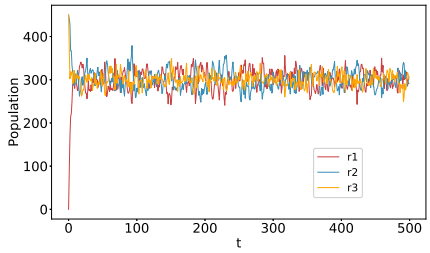


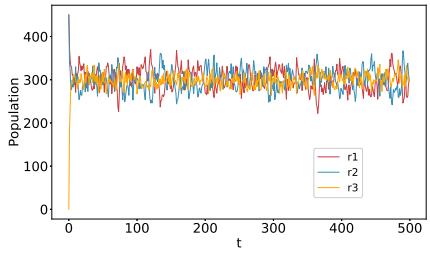
Figure 3.4: Flowcharts of the relations between actions and rules for (a) S_1, (b) S_2, and (c) S_4. The solid arrow represents positive effects from action to rule, and the dashed arrow represents positive effects from rule to action. The width of an arrow indicates how strong the effect is.

In Figure 3.2, we imposed an initial configuration with equal densities of the three rules, but this is not the reason why eventually the three rules have almost equal densities. Figure 3.5 gives four cases of extreme initial configurations, which implies that regardless of the initial densities of the rules, the system will always reach a state where all rules

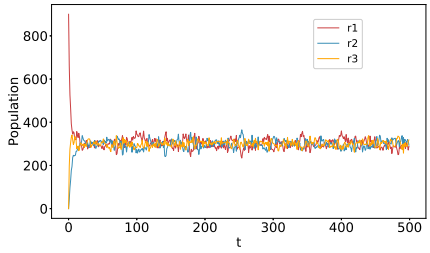
have approximately the same density. On the contrary, the opinion (rule) in the final state of consensus in the voter model is completely determined by the initial configuration. Precisely, the consensus of r_1 occurs with probability $P_1 = p_0$ and the consensus of r_2 occurs with $P_2 = 1 - p_0$, given that the system was initially composed of a fraction p_0 of agents believes in r_1 and a fraction $1 - p_0$ of agents believes in r_2 (Krapivsky et al., 2010). To summarize, the initial configuration determines the result of the voter model, but has no effect on the result of the AOI model with S_4. The different roles of the initial configurations, obviously, result from the emergence/ absence of the inclusive rule r_3. Comparing Figure 3.4 (a) with (c), we can see that r_3 plays the role of a bridge connecting the two opposite pairs (r_1, a_1) and (r_2, a_2), and the bridge helps to balance the densities of rules dynamically. Figure 3.6 illustrates the dynamics in a simple way using the cartoon of a set of communicating vessels.



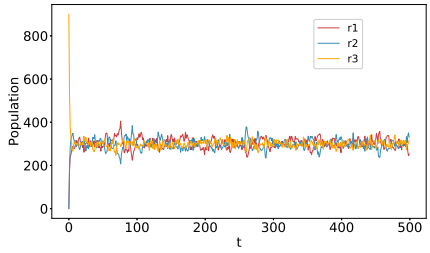
(a) $r_1 p = 0$, $r_2 p = r_3 p = 450$



(b) $r_1 p = r_2 p = 450$, $r_3 p = 0$



(c) $r_1 p = 900$,$r_2 p = r_3 p = 0$



(d) $r_1 p = r_2 p = 0$, $r_3 p = 900$

Figure 3.5: Simulation results of the AOI model with $S = S_4$ on an $L = 30$ lattice with a random initial configuration. $A = \{a_1, a_2\}$, $R = \{r_1, r_2, r_3\}$. The system has an initial population of $r_1 p$ believers in r_1, $r_2 p$ believers in r_2, and $r_3 p$ believers in r_3.

In Figure 3.7 we present the simulation result of the AOI model with S_2. S_2 constitutes another interesting case where a_1 is obliged by r_1 and permitted by r_2, while a_2 is only permitted by a_2. Clearly, a_1 holds a major advantage over a_2, but as we can see in Figure 3.7 (b), a_2 still survives and holds a small but stable fraction of population with the help of r_2. Again the composition of the stable state has no dependence on the initial configurations (Figure 3.8). The flowchart corresponding to S_2 can be found in Figure 3.4 (b).

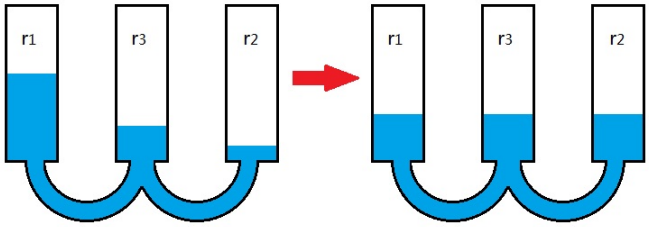
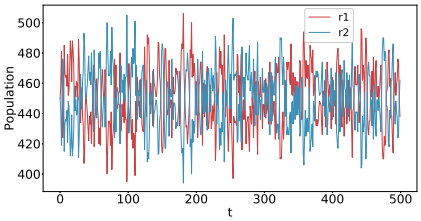
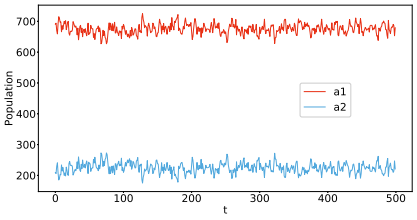


Figure 3.6: Illustration of the AOI model with S_4 in the form of communicating vessels. Containers represent rules, and the liquid level in each container represents the population of the corresponding rule. The left part shows the initial liquid distribution, and the right part shows the stable state of liquid.

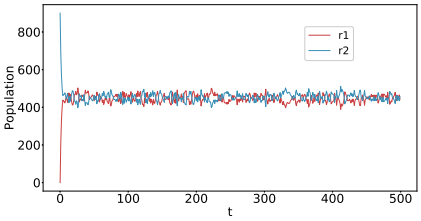


(a) Population distribution of rules

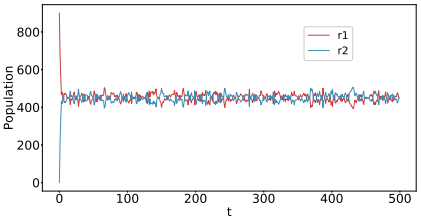


(b) Population distribution of actions

Figure 3.7: Simulation results of the AOI model with $S = S_2$ on an $L = 30$ lattice with a random initial configuration and equal densities of both rules. $A = \{a_1, a_2\}$, $R = \{r_1, r_2\}$.



(a) $r_1 p = 0$, $r_2 p = 900$



(b) $r_1 p = 900$, $r_2 p = 0$

Figure 3.8: Simulation results of the AOI model with $S = S_2$ on an $L = 30$ lattice with a random initial configuration. $A = \{a_1, a_2\}$, $R = \{r_1, r_2\}$. The system has an initial population of $r_1 p$ believers in r_1, and $r_2 p$ believers in r_2.

To summarize, the inclusive rule has three non-trivial effects on opinion dynamics. First, the inclusive rule prohibits the clustering process; second, consensus is never reached if an inclusive rule is present; finally, the composition of the final population is not determined by the initial densities of rules when inclusive rules are present. In other words, the three key features of the voter model, namely clustering, consensus, and

the dependence on the initial configurations, disappear due to the introduction of the inclusive rule. Instead, the inclusive rule leads to a non-clustering dynamics of opinions and a diverse final state of opinions that is not related to the initial configuration at all.

3.4.3. MATHEMATICAL DERIVATION

Given the various forms of action-opinion relations, obtaining a general analytical solution for the AOI model is very difficult, and therefore we have so far used the cellular automata approach to investigating the evolution of the opinions in the previous subsections. However, it is beneficial to provide analytical results for some simple action-opinion matrices, which would help us better understand the evolution dynamics, especially the feature of (in)dependence of initial configurations in a precise manner. In this subsection, we provide the mathematical analysis for S_1 and S_2, whose simulation results have been displayed in Figure 3.1 and Figure 3.7 already. The following derivation generally follows the path that solves the voter model (Krapivsky et al., 2010).

In S_1 or S_2 there are only two rules in total, thus we can define the rules in a binary way: the rule adopted by the agent i, $r^{(i)}$, can be either $+1$ (which means $r^{(i)} = r_1$) or -1 (which means $r^{(i)} = r_2$). We write $r^{(i)}(\tau) = r^{(i)}$ to keep the notation simple. According to equation (3.1) and (3.2), the flip rate that the agent i changes her rule $r^{(i)}$ is:

$$w_i(\mathbf{s}) = \frac{1}{2}\{1 - \frac{r^{(i)}}{4}[\sum_{j\in M_i}\sum_{r} P(r^{(j)} = r|a^{(j)})r]\} \tag{3.3}$$

where $\mathbf{s}$ is the current configuration of the system, and the subscript i in $w_i(\mathbf{s})$ implies that only agent i changes her rule in an update (i.e., a time unit, see Section 3.3 for reference). The scalar r equals either $+1$ or -1. The flip rate shown in equation (3.3) is analogous to the flip rate in the voter models (Krapivsky et al., 2010).

The master equation is easy to derive, but difficult to solve. Instead, we focus on the average opinion (rule) for each agent, namely $R(i) \equiv < r^{(i)} >$, where $< \cdot >$ is the average notation, defined by $< x >= \sum_{x'} P(x = x')x'$. In a short enough time interval $\Delta\tau$, the rule of agent i changes according to:

$$r^{(i)}(\tau + \Delta\tau) = \begin{cases} r^{(i)}(\tau) & \text{with probability } 1 - w_i(\mathbf{s})\Delta\tau \\ -r^{(i)}(\tau) & \text{with probability } w_i(\mathbf{s})\Delta\tau \end{cases} \tag{3.4}$$

Following Krapivsky's path (Krapivsky et al., 2010), from (3.4) we notice that agent i's opinion changes by $-2r^{(i)}$ with an instant probability $w_i(\mathbf{s})$, then the evolution dynamics of the average opinion is:

$$\frac{dR(i)}{d\tau} = \frac{d < r^{(i)} >}{d\tau} = lim_{\Delta\tau\to 0} < \frac{r^{(i)}(\tau + \Delta\tau) - r^{(i)}(\tau)}{\Delta\tau} >= -2 < r^{(i)} w_i(\mathbf{s}) > \tag{3.5}$$

Substitute equation (3.3) into (3.5) and use $(r^{(i)})^2 = 1$:

$$\frac{dR(i)}{d\tau} = - < r^{(i)}\{1 - \frac{r^{(i)}}{4}[\sum_{j\in M_i} P(r^{(j)} = r|a^{(j)})r\}] >= -R(i) + \frac{1}{4}\sum_{j\in M_i} < \sum_{r} P(r^{(j)} = r|a^{(j)})r > \tag{3.6}$$

and define $< \sum_r P(r^{(j)} = r|a^{(j)})r > \equiv R^\star(j)$, which is agent i's perceived average opinion of agent j, gives

$$\frac{dR(i)}{d\tau} = -R(i) + \frac{1}{4}\sum_{j\in M_j} R^\star(j) \tag{3.7}$$

In the voter model, $R(j) = R^\star(j)$, so the equation reduces to $\frac{dR(i)}{d\tau} = -R(i) + \frac{1}{4}\sum_{j\in M_j} R(j)$. Analogous to magnetization in the vote model, we define the mean magnetization of the system as: $m \equiv \sum_i R(i)/N$, which measures the average opinion of the whole system, and $m = +1$ means the system reaches the consensus of r_1, while $m = -1$ means the consensus of r_2. Summing equation (3.7) over all agents:

$$N\frac{dm}{d\tau} = -\sum_i R(i) + \frac{1}{4}\sum_i \sum_{j\in M_i} R^\star(j) \tag{3.8}$$

If we take a close look at $R^\star(j)$, since r can be either $+1$ or -1, we can rewrite $R^\star(j)$ as:

$$R^\star(j) \equiv < \sum_r P(r^{(j)} = r|a^{(j)})r > = < P(r^{(j)} = 1|a^{(j)}) - P(r^{(j)} = -1|a^{(j)}) > = 2 < P(r^{(j)} = 1|a^{(j)}) > -1 \tag{3.9}$$

where we've used $P(r^{(j)} = 1|a^{(j)}) + P(r^{(j)} = -1|a^{(j)}) \equiv 1$. Now by using the action-opinion matrix, we can solve for $< P(r^{(j)} = 1|a^{(j)}) >$.

[S_1] From S_1, we know:

$$P(r^{(j)} = 1|a^{(j)}) = \begin{cases} 1 & \text{with probability } P(a^{(j)} = a_1) \\ 0 & \text{with probability } P(a^{(j)} = a_2) \end{cases} \tag{3.10}$$

Therefore

$$< P(r^{(j)} = 1|a^{(j)}) > = P(a^{(j)} = a_1) \tag{3.11}$$

On the other hand, one can rewrite

$$R(i) = < r^{(i)} > = P(r^{(i)} = 1) - P(r^{(i)} = -1) = 2P(r^{(i)} = 1) - 1 \tag{3.12}$$

Substitute equation (3.9), (3.11) and (3.12) into equation (3.8):

$$N\frac{dm}{d\tau} = -\sum_i (2P(r^{(i)} = 1) - 1) + \frac{1}{4}\sum_i \sum_{j\in M_i} [2P(a^{(j)} = a_1) - 1] \tag{3.13}$$

Rearranging and simplifying (3.13) gives:

$$N\frac{dm}{d\tau} = 2\sum_i [P(a^{(i)} = a_1) - P(r^{(i)} = 1)] \tag{3.14}$$

where we have used the trick that $\sum_i \sum_{j\in M_i} [2P(a^{(j)} = a_1) - 1] = 4\sum_i [2P(a^{(i)} = a_1) - 1]$.

For S_1, it is clear that $P(a^{(i)} = a_1) = P(r^{(i)} = 1)$ because believing in r_1 is equivalent to acting as a_1, and vice versa. Therefore we have $N\frac{dm}{d\tau} = 0$, which means the magnetization m is conserved in the AOI model with S_1 (voter model). The conserved magnetization helps to understand the features of the voter model stated in Section 3.4.1. Also, the result is identical to the result solved for the voter model.

[S_2] From S_2, it can be calculated that:

$$P(r^{(j)}=1|a^{(j)})=\begin{cases}\frac{2}{3} & \text{with probability } P(a^{(j)}=a_1)\\ 0 & \text{with probability } P(a^{(j)}=a_2)\end{cases} \tag{3.15}$$

Therefore

$$<P(r^{(j)}=1|a^{(j)})>=\frac{2}{3}P(a^{(j)}=a_1) \tag{3.16}$$

Substitute equation (3.9), (3.12) and (3.16) into equation (3.8) and simplify it:

$$N\frac{dm}{d\tau}=2\sum_i[\frac{2}{3}P(a^{(i)}=a_1)-P(r^{(i)}=1)] \tag{3.17}$$

By conditional probability calculus, we can obtain that:

$$P(a^{(i)}=a_1)=P(r^{(i)}=1)P(a_1|r^{(i)}=1)+P(r^{(i)}=-1)P(a_1|r^{(i)}=-1) \tag{3.18}$$

S_4 shows that $P(a_1|r^{(i)}=1)=1$ and $P(a_1|r^{(i)}=-1)=0.5$, thus (3.18) becomes:

$$P(a^{(i)}=a_1)=P(r^{(i)}=1)+\frac{1}{2}P(r^{(i)}=-1)=\frac{1}{2}P(r^{(i)}=1)+\frac{1}{2} \tag{3.19}$$

where we have used $P(r^{(i)}=1)+P(r^{(i)}=-1)=1$. Substitute equation (3.19) into equation (3.17) and rearrange it:

$$N\frac{dm}{d\tau}=2\sum_i\{\frac{1}{3}-\frac{2}{3}P(r^{(i)}=1)\}=\frac{2}{3}N-\frac{4}{3}\sum_i P(r^{(i)}=1) \tag{3.20}$$

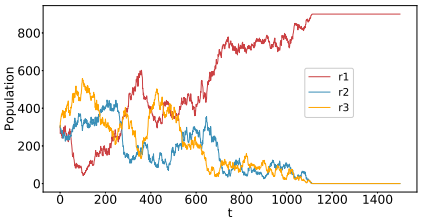
Thus the fixed point is $\sum_i P(r^{(i)}=1)=N/2$, that is, the probability of believing in r_1, averaged over the population, is 1/2. Starting from any configuration that $\sum_i P(r^{(i)}=1)>N/2$, for example, the configuration where $r^{(i)}=1$, $\forall i$, since $N\frac{dm}{d\tau}<0$, will always converge to a (dynamic) state where $\sum_i P(r^{(i)}=1)=N/2$, which is the stable state shown in Figure 3.7 (a). Similarly, the system starting with the configuration where $\sum_i P(r^{(i)}=1)<N/2$, because $N\frac{dm}{d\tau}>0$, will still evolve to the same state where $\sum_i P(r^{(i)}=1)=N/2$. The analytical result helps us understand that in the model with S_2, why the composition of the stable state is independent of the initial configuration (Figure 3.8), and why the mixed state of rules is always the final stable state (Figure 3.7 (a)).

3.5. THREE-ACTION SITUATION

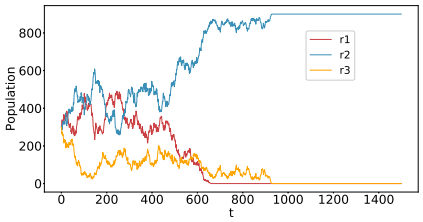
Although most studies only deal with two-state voter model due to simplicity, it is promising to study the AOI model with three actions because a larger number, and more subtle, action-opinion relations are possible compared to the two-action situation. For reasons of space limitations, we will not go through all the possible situations, but focus on the cases shown by S_5, S_6 and S_7 specifically:

$$S_5=\begin{array}{c}\\ r_1\\ r_2\\ r_3\end{array}\begin{array}{c}\begin{array}{ccc}a_1 & a_2 & a_3\end{array}\\ \begin{pmatrix}+ & - & -\\ - & + & -\\ - & - & +\end{pmatrix}\end{array}\qquad S_6=\begin{array}{c}\\ r_1\\ r_2\\ r_3\\ r_4\end{array}\begin{array}{c}\begin{array}{ccc}a_1 & a_2 & a_3\end{array}\\ \begin{pmatrix}+ & - & -\\ - & + & -\\ - & - & +\\ 0 & 0 & -\end{pmatrix}\end{array}\qquad S_7=\begin{array}{c}\\ r_1\\ r_2\\ r_3\\ r_4\end{array}\begin{array}{c}\begin{array}{ccc}a_1 & a_2 & a_3\end{array}\\ \begin{pmatrix}+ & - & -\\ - & + & -\\ - & - & +\\ 0 & 0 & 0\end{pmatrix}\end{array}$$

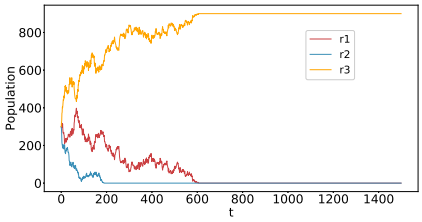
The AOI model using S_6 is nothing but a three-state voter model, which is widely used in the studies of language competition (Castelló et al., 2006; Hadzibeganovic et al., 2008). Unsurprisingly, all features of the two-state voter model (Figure 3.1) are still valid in the three-state voter model: Figure 3.9 illustrates that consensus is always reached, and each type of consensus (i.e., Figure 3.9 (a), (b), and (c)) has the same probability to become the final absorbing state because the initial densities of all rules are set equal. Additionally, the clustering phenomenon of the three rules is shown in Figure 3.10.



(a) Consensus of r_1



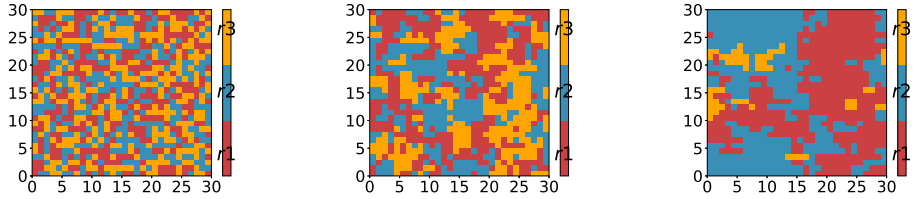
(b) Consensus of r_2



(c) Consensus of r_3

Figure 3.9: Three possible simulation results of the AOI model with $S = S_5$ on an $L = 30$ lattice with a random initial configuration and equal density of each rule. $A = \{a_1, a_2, a_3\}$, $R = \{r_1, r_2, r_3\}$.

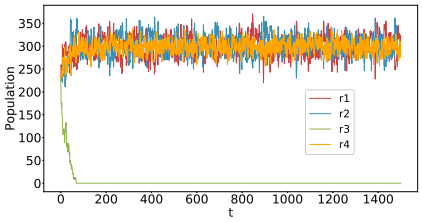
S_6 is constructed by introducing an inclusive rule r_4 (that permits a_1 and a_2 but prohibits a_3) to S_5, so the difference between the simulations result of S_5 and S_6 implies the role of what we call a **preferentially inclusive rule**. Rule r_4 in S_6 is called a preferentially inclusive rule because it shows strict preference for a_1 and a_2 over a_3, although it is indifferent between a_1 and a_2. On the other hand, r_2 in S_2, r_3 in S_4, and r_4 in S_7 are called **non-preferentially inclusive rules** because they are completely indifferent to any action.



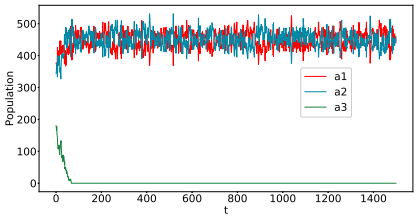
(a) $t = 1$ (b) $t = 5$ (c) $t = 100$

(d) $t = 600$ (e) $t = 800$ (f) $t = 1000$

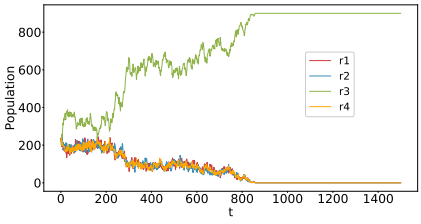
Figure 3.10: Snapshots of rule distribution of the AOI model presented in Figure 3.9 (b).



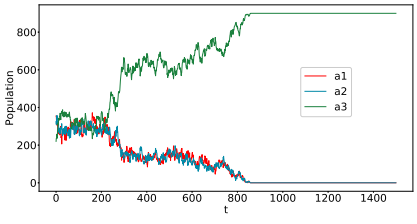
(a) case A: dynamics of rules



(b) case A: dynamics of actions

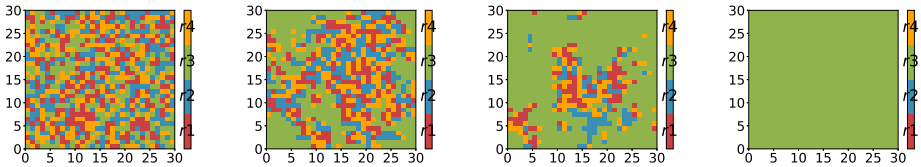


(c) case B: dynamics of rules



(d) case B: dynamics of actions

Figure 3.11: Two cases of simulation results of the AOI model with S_6 on an $L = 30$ lattice with a random initial configuration and equal density of each rule. $A = \{a_1, a_2, a_3\}$, $R = \{r_1, r_2, r_3, r_4\}$.



(a) $t = 1$ (b) $t = 200$ (c) $t = 400$ (d) $t = 1000$

Figure 3.12: Snapshots of the rule distribution of the AOI model presented in Figure 3.11 (c).

It is reported that there are two types of outcomes given the action-opinion matrix S_6, namely case A (Figure 3.11 (a) & (b)) and case B (Figure 3.11 (c) & (d)). In case A, r_4 dies out rapidly, while r_1, r_2, and r_3 coexist. Specifically, one of r_1 and r_2 is in a momentary majority alternatively, but r_4 holds a relatively stable share over time (Figure 3.11 (a)). The dynamics of rules lead to similar evolutionary paths for actions (Figure 3.11 (b)), where a_3 go extinct and a_1 and a_2 coexist. In case B, consensus of r_3 is reached eventually, and r_1, r_2 and r_4 die out gradually. One can not predict which case we will obtain from one realization of the simulation. There relation between case A and B mimics a trade-off between a smaller chance to become the sole superpower that dominates everyone (r_3 in case B), and a larger chance to dominate around 1/3 of the population (r_1, r_2 and r_4 in case A). In 100 independent trials, we find that **74%** trials are in case A, and **26%** trials are in case B. The distribution of case A and B suggests that the trade-off is in equilibrium: consider a finite system with equal density for each rule r_1, r_2, r_3 and r_4. Ultimately, the system reaches the consensus of r_3 with probability P_B, and reaches the mixed state of r_1, r_2, and r_4 with probability P_A. So the expected number of agents believes in r_3, $< N_3 >$, should be $P_B N$, and the the expected number of believes in any other rule, $< N_k >$ ($k = 1, 2, 4$), is $\frac{1}{3} P_A N$. In our trials, we observed that 74 trials are in case A and 26 in case B, so the estimated P_A, $\hat{P_A}$, is 0.74, and the estimated P_B, $\hat{P}_B$ is 0.26. Substitute the two estimated probabilities, we find that:

$$\frac{1}{3} \hat{P}_A N \approx \hat{P}_B N \tag{3.21}$$

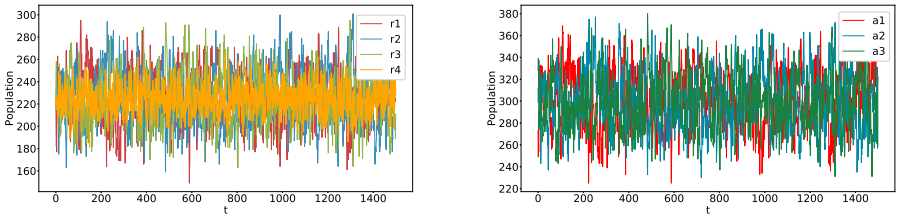
Based on the trials, we conjecture that $P_A = 0.75$, and $P_B = 0.25$, which leads to:

$$< \hat{N}_1 > = < \hat{N}_2 > = < \hat{N}_3 > = < \hat{N}_4 > \tag{3.22}$$

which implies that all rules have the same expected population of believers. Consequently, the trade-off is in equilibrium. To better understand the result, imagine a gamble where the player is asked to bet on the most popular rule in the AOI model described by S_6. Equation (3.22) tells her that she should be indifferent to any choice, as all betting strategies lead to the same expected payoff.

The evolution of case B is illustrated in Figure 3.12, describing how the system reaches the consensus of r_3 from a mixed state of all rules. An interesting observation is that besides the single-rule clusters of r_3, there are also some mixed-rule clusters composed of r_1, r_2 and r_4. Comparing Figure 3.12 with Figure 3.10, we can see that the preferentially inclusive rule r_4 reduces the ability of r_1 or r_2 to form a single-rule cluster of its own. The

results of the AOI model with S_6 shows that there are two categories for exclusive rules: r_1 and r_2 form a category that coexist with r_4 and cannot form single-rule clusters, and r_3 itself is another category, as r_3 and r_4 are completely incompatible. Here we define that two rules are **compatible** with each other if there exists at least one action that is allowed (i.e., *obliged* or *permitted*) by both of them, and otherwise we say that they are **incompatible**. If we look at S_6 carefully, it is clear that r_1 and r_2 are compatible with r_4 respectively: both r_1 and r_4 allow a_1, and both r_2 and r_4 allow a_2. On the contrary, the two actions permitted by r_4 (i.e., a_1 and a_2) are prohibited by r_3, and the only action obliged by r_3 (i.e., a_3) is forbidden by r_4, so r_3 and r_4 are incompatible. The different relations between exclusive rules and inclusive rules explain the different clustering features of each rule. Because both r_1 and r_2 are compatible with r_3 , the three rules coexist and no single-rule clusters can be formed. Meanwhile, r_3 and r_4 are incompatible, so r_3 is unlikely to coexist with r_4. Since r_1 and r_2 coexist with r_4, r_3 cannot coexist with r_1 and r_2 either. Thus the single-rule cluster of r_3 emerges if it dominates the population (i.e., case B in Figure 3.11).



(a) $t = 1$ (b) $t = 200$

Figure 3.13: Simulation results of the AOI model with $S = S_7$ on an $L = 30$ lattice with a random initial configuration and equal density of each rule. $A = \{a_1, a_2, a_3\}$, $R = \{r_1, r_2, r_3, r_4\}$.

The only difference between S_6 and S_7 lies in the inclusive rule. In S_6, r_4 is a preferentially inclusive rule that permits a_1 and a_2 but prohibits a_3. However, in S_7 we are having a non-preferentially inclusive rule r_4, which permits all actions. Given S_7, now all exclusive rules and the inclusive rule are compatible, so all rules can coexist (Figure 3.13). It should be noted that although all rules share almost the same fraction of population on average, the variation of the population believing in r_4 is significantly smaller than other rules. It is equivalent to saying that the share of the population which employs exclusive rules is more likely to be either very small or very large, while the share of the population which employs the inclusive rule is of intermediate magnitude and relatively stable through time.

To summarize, the inclusive rule "forms an alliance" with all the exclusive rules that are compatible with it (e.g., r_4 and r_1, r_2 in S_6) to compete with, if it exists, the exclusive rule with which it is incompatible (e.g., r_3 in S_6). In S_6 r_1, r_2 and r_4 form an alliance against r_4, while in S_7 all rules constitute a large alliance. In both cases, the expected population of the believers in each rule in the alliance should be the same (Figure 3.11 & 3.13).

3.6. DISCUSSION AND CONCLUSION

3.6.1. DISCUSSION: CONSTRAINED VOTER MODEL, LANGUAGE COMPETITION, AND THE AOI MODEL

The above simulations have proved that the action-opinion matrix is the most important factor determining the results obtained from the AOI model. A question that arises naturally is where the matrix originates from? Actually, this question concerns the relations between actions, something which we have not discussed in depth yet. In S_1, a_1 and a_2 are two excluding actions, in the sense that it is impossible to be indifferent between a_1 and a_2, and agents must have a strict preference. On the contrary, in S_4, r_3 offers an option for centrists (i.e., the believers in inclusive rules): believers in r_3 are indifferent between a_1 and a_2, and thus they choose actions randomly. This is related to the constrained voter model (Vazquez et al., 2003) and its modification (de la Lama et al., 2006) where agents can be extremists (including leftists and rightists) or centrists. In those models, the centrists, or the undecided agents, serve as an intermediate group that can be converted to one of the extremists, while the extremists do not interact with each other (Castellano et al., 2009). This is also a common method to implement a bilingual state in language competitions, where any change between the two monolingual states must go through an intermediate state called the bilingual state (Castelló et al., 2006; Colaiori et al., 2015).

Although having different formats, the inclusive rule in the AOI model, the centrists in the constrained voter model, and the bilinguals in the language competition model all describe an intermediate state that bridges the other two (or possibly more than two in the AOI model) non-excluding states. Pioneering researchers have already acknowledged the existence of an intermediate state, but the absence of action-opinion relations limits the scope: the intermediate state must include all existing states in the constrained models and the language competition model (but not in the AOI model). To summarize, the two-state classic voter model resembles the AOI model with S_1, while the constrained voter model and the bilingual language competition model resemble the AOI model with S_4. Obviously, the AOI model provides more possibilities other than S_1 and S_4 by introducing various action-opinion relations. Finally, it is worth mentioning that although the constrained voter model and the bilingual language competition model are similar to the AOI model with S_4, due to the different dynamical rules to update agent's opinion, the three models lead to completely different results. In the constrained voter model, the final states are a consensus in one of the three states or a mixture of the extremists. The bilingual language competition model always ends up with one of the monolingual consensuses (Castellano et al., 2009). Conversely, the AOI model with S_4, as stated in Figure 3.2, provides another outcome where the mixed state of all states: r_1 (resembles one of the extremists or monolingual states), r_2 (the same resemblance as r_1), and r_3 (resembles the centrists, or the bilingual state), which is an impossible outcome for the other two models.

3.6.2. BRIEF CONCLUSION AND OUTLOOK

The most important contribution of this paper is to provide an alternative and – in our view – more realistic approach to modeling the spreading of opinions compared to existing models of opinion dynamics. The new approach, called the Action-Opinion Inference

(AOI) model, is based on the postulate that opinions themselves are unobservable, but may be learned by observing the actions that are governed by the opinion; this learning process may be partial given that actions are noisy signals of underlying opinions due to the multiplicity of action-opinion relations. The AOI model captures the *"learning opinions by observing actions"* process, which is an intuitive assumption but has been ignored in the studies of opinion dynamics. In the AOI model, an agent first observes the actions of her neighbors, and then infers her neighbors' opinions (represented by rules) according to the observations. Then the agent updates her own rule based on the perceived probabilities of each rule among her neighbors.

We show that the outcome of the AOI model strongly depends on action-opinion relations, described by the action-opinion matrix. When the mapping of the action set A to the rule set R (or vice versa) is a bijective function (e.g., S_1 and S_5), the AOI model reduces to a classic voter model. When introducing an inclusive rule that permits all actions to the bijective relation (e.g., S_4), the model resembles the constrained voter model. The variation of the action-opinion matrix enables us to investigate a broad range of opinion dynamics. A striking finding from the simulation results for the two-action AOI model is the role of inclusive rules, defined as the rules that permit more than one action, in a competition with other rules. An inclusive rule bridges the actions it permits, which means the rule also bridges the exclusive rules that oblige these actions. An exclusive rule is defined as a rule that obliges only one action. The connection between exclusive rules via the inclusive rule(s) leads to a final mixed state of all these rules, regardless of the initial density of each rule. This phenomenon has never been found in either constrained voter models or bilingual language competitions. The three-action AOI model is more complex, where the inclusive rule forms an alliance with all the exclusive rules that share at least one allowed action to compete with the exclusive rule(s) that shares no allowed actions with the inclusive rule (i.e., S_6). The competition between the alliance and the incompatible exclusive rule is a winner-take-all game, but if the alliance wins the whole population, the members share the population equally on average.

Admittedly, the major limitation of the AOI model is the difficulty to incorporate complex reality into a simple action-opinion matrix. As mentioned before, the simulation outcomes, as well as the analytical solutions, are based on the relations between actions and opinions, which in this paper are represented by a series of simple action-opinion matrices. In fact, the reality is far more complicated than all the matrices we have shown in the paper. First, we are not sure how many underlying opinions people can infer from an action, and sometimes multiple opinions collectively lead to one action. Taking the example of cycling again, it is unfeasible to list all the possible opinions that lead to the action of cycling: besides being environmentally-friendly or cost-sensitive, the cyclist may simply love this sport, or actually he just randomly chooses a travel mode and it happens to be cycling today. Moreover, we are not sure if every agent in the society is aware of all the possible opinions. With a slightly different matrix, the simulation result could be different. Given this limitation, we would recommend first applying the model to some simple and obvious situations. Second, further modifications to the design of action-opinion matrices can be a major challenge in future work to enhance the model's capability to describe reality. It should also be noted that changing the current assumption of discrete opinions, described by "+", "−" and "0", to a more realistic but complicated

assumption of cardinal opinions might fit the reality better. Cardinal opinions offer more ways to describe evaluations, rather than simply referring to an action being completely forbidden and completely obliged (or permitted); a consequence would be that a matrix of finite size would no longer represent the full set of possible action-opinion relations. Given that the central concept of the uncertainty in the relations between actions and opinions, which is the key to explaining simulation outcome's independence of initial configurations (see Section 3.4 and Section 3.5), remains unchanged, we optimistically speculate that the main result of the discrete opinion version should be robust in the cardinal opinion version of the model.

In all, the AOI model establishes a new framework for researchers to cope with the latency of opinions and with a variety of presumed action-opinion relations. We believe that the AOI model does not only serve as another modification of the voter model, but also constitutes an attempt to study the spreading of both actions and opinions while opening the floor for further discussions in opinion dynamics. Despite the fact that there are still some possible action-opinion matrices that we have not tested yet in the three-action situation, several avenues for further research are promising. First, the AOI model can be extended or adapted by employing other methods that represent processes of "learning opinions by observing actions" to make the model more realistic. For instance, a similarity-based mechanism may assume that an agent is more likely to take the opinion that is similar to her previous opinion (Teşileanu & Meyer-Ortmanns, 2006; Flache et al., 2017). In addition, it would be interesting to explore opinion dynamics in the situation where (some) agents are reluctant to signal their opinions through their actions. Such obfuscation behavior, which is characterized by an agent choosing an action that provides minimal information to a focal agent regarding her underlying opinion, has been formalized in recent work (Chorus, 2018). Second, the model can be tested in various network structures. We only test the model in the von Neumann neighborhood in the paper, and it is promising to analyze the dynamics of the AOI model in different networks to investigate the role of randomness, degree distribution, and dimensionality. Furthermore, analytical solutions to the model with general action-opinion relations (a simple example has been given in Section 3.4) would be helpful to understand the simulation results. Finally, an obvious and important direction for further research consists of empirically validating – at a micro and macro level – our behavioral model and the emergent properties it generates.

ACKNOWLEDGMENTS

This research has received funding from the European Research Council: Consolidator Grant BEHAVE (grant agreement No. 724431). A prior version of this research has been presented at the 21st International Conference on Principles and Practice of Multi-Agent Systems (PRIMA 2018). T. Tang would like to acknowledge the helpful comments from session chair Michael Mäs and other participants of the conference. The authors would also like to thank the anonymous referee for helpful suggestions.

REFERENCES

[1] Axelrod, R. (1997). The dissemination of culture: A model with local convergence and global polarization. Journal of Conflict Resolution, 41(2), 203-226.

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

[3] Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. Journal of Political Economy, 100(5), 992-1026.

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

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

[6] Castelló, X., Eguíluz, V. M., & San Miguel, M. (2006). Ordering dynamics with two non-excluding options: bilingualism in language competition. New Journal of Physics, 8(12), 308.

[7] Chorus, C. G. (2018). A simple model of obfuscation-based decision-making by human and artificial agents. Working Paper.

[8] Colaiori, F., Castellano, C., Cuskley, C. F., Loreto, V., Pugliese, M., & Tria, F. (2015). General three-state model with biased population replacement: Analytical solution and application to language dynamics. Physical Review E, 91(1), 012808.

[9] de la Lama, M. S., Szendro, I. G., Iglesias, J. R., & Wio, H. S. (2006). Van Kampen's expansion approach in an opinion formation model. The European Physical Journal B-Condensed Matter and Complex Systems, 51(3), 435-442.

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

[11] Dornic, I., Chaté, H., Chave, J., & Hinrichsen, H. (2001). Critical coarsening without surface tension: The universality class of the voter model. Physical Review Letters, 87(4), 045701.

[12] Fishbein, M. (1963). An investigation of the relationships between beliefs about an object and the attitude toward that object. Human Relations, 16(3), 233-239.

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

[14] 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 & Social Simulation, 20(4).

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

[16] Hadzibeganovic, T., Stauffer, D., & Schulze, C. (2008). Boundary effects in a three-state modified voter model for languages. Physica A: Statistical Mechanics and its Applications, 387(13), 3242-3252.

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

[18] 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.

[19] 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).

[20] Krapivsky, P. L., & Redner, S. (2003). Dynamics of majority rule in two-state interacting spin systems. Physical Review Letters, 90(23), 238701.

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

[22] Lambiotte, R., & Redner, S. (2007). Dynamics of vacillating voters. Journal of Statistical Mechanics: Theory and Experiment, 2007(10), L10001.

[23] Latané, B. (1981). The psychology of social impact. American Psychologist, 36(4), 343–356.

[24] Liggett, T. M. (1994). Coexistence in threshold voter models. The Annals of Probability, 22(2), 764-802.

[25] Liska, A. E. (1984). A critical examination of the causal structure of the Fishbein/Ajzen attitude-behavior model. Social Psychology Quarterly, 47(1), 61-74.

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

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

[28] Mäs, M., & Flache, A. (2013). Differentiation without distancing. Explaining opinion bipolarization without assuming negative influence. PLoS ONE, 8(11), e74516.

[29] 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.

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

[31] Mobilia, M. (2003). Does a single zealot affect an infinite group of voters?. Physical Review Letters, 91(2), 028701.

[32] Moons, I., & De Pelsmacker, P. (2012). Emotions as determinants of electric car usage intention. Journal of Marketing Management, 28(3-4), 195-237.

[33] Pacheco, J. (2012). The social contagion model: Exploring the role of public opinion on the diffusion of antismoking legislation across the American states. The Journal of Politics, 74(1), 187-202.

[34] Petty, R. E., Cacioppo, J. T., & Goldman, R. (1981). Personal involvement as a determinant of argument-based persuasion. Journal of Personality and Social Psychology, 41(5), 847-855.

[35] Prelec, D. (2004). A Bayesian truth serum for subjective data. Science, 306(5695), 462-466.

[36] Seeme, F. B., & Green, D. G. (2016). Pluralistic ignorance: Emergence and hypotheses testing in a multi-agent system. In 2016 International Joint Conference on Neural Networks (IJCNN) (pp. 5269-5274).

[37] Sood, V., & Redner, S. (2005). Voter model on heterogeneous graphs. Physical Review Letters, 94(17), 178701.

[38] Suchecki, K., Eguíluz, V. M., & San Miguel, M. (2005). Voter model dynamics in complex networks: Role of dimensionality, disorder, and degree distribution. Physical Review E, 72(3), 036132.

[39] Svenkeson, A., & Swami, A. (2015). Reaching consensus by allowing moments of indecision. Scientific Reports, 5, 14839.

[40] Sznajd-Weron, K. (2005). Sznajd Model and its applications. Acta Physica Polonica B, 36(8), 2537-2547.

[41] Teşileanu, T., & Meyer-Ortmanns, H. (2006). Competition of languages and their hamming distance. International Journal of Modern Physics C, 17(02), 259-278.

[42] Valente, T. W. (1996). Social network thresholds in the diffusion of innovations. Social Networks, 18(1), 69-89.

[43] Vazquez, F., Krapivsky, P. L., & Redner, S. (2003). Constrained opinion dynamics: Freezing and slow evolution. Journal of Physics A: Mathematical and General, 36(3), L61.

[44] Wang, Z., Liu, Y., Wang, L., & Zhang, Y. (2014). Freezing period strongly impacts the emergence of a global consensus in the voter model. Scientific Reports, 4, 3597.

[45] Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. Nature, 393(6684), 440-442.

[46] Wicker, A. W. (1969). Attitudes versus actions: The relationship of verbal and overt behavioral responses to attitude objects. Journal of Social Issues, 25(4), 41-78.

[47] Wu, F., & Huberman, B. A. (2004). Social structure and opinion formation. arXiv preprint cond-mat/0407252.

[48] Zanna, M. P., Olson, J. M., & Fazio, R. H. (1980). Attitude–behavior consistency: An individual difference perspective. Journal of Personality and Social Psychology, 38(3), 432-440.

4

HIDING OPINIONS BY MINIMIZING DISCLOSED INFORMATION: AN OBFUSCATION-BASED OPINION DYNAMICS MODEL

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

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

Keywords: Obfuscation, Opinion-hiding, Opinion dynamics, Agent-based modeling, Social simulation

4.1. INTRODUCTION

For diverse reasons, people may be unwilling to disclose their opinions to the public, especially when the topic is controversial. Instead, they may try to hide their opinions by

adopting various strategies. As a prevalent behavior, hiding opinions has been studied in a number of opinion dynamics models. The majority of them assume that individuals hide their opinions by simply keeping silent (e.g., Gawronski et al., 2014; Ross et al., 2019; Sohn, 2019; Sohn & Geidner, 2016; Takeuchi et al., 2015). Although silence may help hide opinions from hearers of our words, it may not hide them from observers of our actions. That is, we usually learn others' opinions by inferring them from their actions based on some universal knowledge about how opinions and actions relate (Tang & Chorus, 2019). In other words, in daily life, actions are known to signal opinions, and as such, keeping completely silent is no longer feasible when observers learn opinions by observing actions[1]. For example, suppose a group of friends containing both vegetarians and omnivores dine in a restaurant where only two dishes are available: beef steak and vegetable salad. If an omnivore wants to hide the opinion that it is OK to eat meat, choosing steak is certainly a bad idea, but keeping silent is also impractical. A better strategy is to choose salad, because both vegetarians and omnivores are more or less likely to eat salad, but only omnivores will order steak. Such a strategy, where actions are chosen that provide minimal information about underlying beliefs and preferences, is called obfuscation; it can be conceived as a manner to minimize the information disclosed to others by producing ambiguity and uncertainty (Chorus et al., 2021). Obfuscation has been a popular topic in software engineering (You & Yim, 2010) and more recently in privacy research (Brunton & Nissenbaum, 2015), but has not yet attracted attention in the community of opinion dynamics. When obfuscation behaviors are prevalent, a failure to capture them in models of opinion dynamics could lead to a biased understanding of how hiding opinions affects public opinion formation.

In this paper, we present an obfuscation-based opinion dynamics model to study the role of obfuscation in public opinion formation by embedding the obfuscation mechanism (Chorus et al., 2021) into the Action-Opinion Inference (AOI) modeling framework (Tang & Chorus, 2019), where people choose actions according to their opinions and learn others' opinions by interpreting their actions. Within this AOI framework, an obfuscating individual would hide her opinion by choosing the action that (i) is permitted by or in line with her opinion, yet (ii) releases the least amount of information about the opinion to others. This model fills the gap between existing models where hiding opinions equates to keeping silent and the reality that the mechanism of hiding opinions can be more subtle and complex than simply staying silent. As a result, our model can offer a more realistic and reasonable explanation of various social phenomena related to public opinion formation, particularly on (morally) sensitive topics. For example, incorporating obfuscation in the model can lead people to overestimate the popularity of the opinion that obliges the observed action. In a relatively simple setting, this may result in a significantly larger population believing in this opinion, which would have otherwise been different if only "silent-keeping" was considered.

The remainder of the paper is organized as follows: In Section 4.2, we review existing opinion dynamics models of hiding opinions and explain the foundation of our model. Section 4.3 describes the model in detail. In Section 4.4, we provide two illustrative examples abstracted from daily life and tales, in order to illustrate how this model works.

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

Section 4.5 provides a brief summary and outlooks for further research.

4.2. THEORETICAL BACKGROUND

4.2.1. HIDING OPINIONS IN OPINION DYNAMICS

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

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

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

4.2.2. OBFUSCATION AND ACTION-OPINION INFERENCE

Opinions are not always expressed by words but can also be revealed by actions. As argued in Section 4.1, when observers learn someone's opinion by observing her actions, keeping silent is not (always) possible; we claim that in such a case, obfuscation becomes the best

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

strategy.

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

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

In recent years, the concept of obfuscation has been introduced and formalized as a communication strategy in choice modeling. Chorus et al. (2021) combine the notions of Bayesian inference and Shannon entropy, integrating them into a formal model of obfuscation-based decision-making. The idea is that a subject knows that her actions signal her underlying preferences (opinions) and selects the action that is in line with her preferences while providing as little as possible information to observers. The model is designed to describe the behaviors of humans whose actions are observed by others as well as the behaviors of autonomous agents under the surveillance of a human supervisor. In the model, agents choose actions based on a particular rule (here: opinion) that is unknown to the supervisor. Based on the observation of the agent's action, the supervisor infers the opinion that motivates the action according to the Bayes' Theorem. An obfuscating human or autonomous agent, being aware that it might be "punished" if the observer or supervisor learns that it has an "unwanted" opinion, will choose actions by maximizing the Shannon entropy generated by its choice while staying as close as possible to its opinion.

To utilize this mechanism in the context of opinion dynamics, we first need to formalize how opinions are learned by observing actions. In fact, such a formalization exists in the form of a so-called Action-Opinion Inference (AOI) model (Tang & Chorus, 2019). In the AOI model, the relation between opinions and actions is described by deontic logic: an opinion can oblige, permit, or prohibit an action. Equipped with the action-opinion relation, individuals "infer the opinions of others by observing and interpreting their actions" (Tang & Chorus, 2019). Based on the inference, individuals update their own opinions "according to the relative probability of each opinion in the neighborhood, cal-

culated from the inferences of different opinions" (Tang & Chorus, 2019). As the final step, individuals choose new actions according to the newly updated opinions. The AOI model is compatible with Chorus et al.'s obfuscation mechanism not only because it formalizes the notion of "learning opinions by observing actions", but also because of the deontic logic underlying the action-opinion relation, where an action may be driven by different opinions, and an opinion may permit different actions, allowing agents to obfuscate by choosing certain actions. If each action is driven by only one opinion, observers can then directly and correctly read opinions from actions, and there will be no room for obfuscation.

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

4.3. THE MODEL

In this section, we develop an opinion dynamics model of obfuscation by embedding the obfuscation mechanism (Chorus et al., 2021) in the framework of the Action-Opinion Inference (AOI) model (Tang & Chorus, 2019).

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

Assume that there is a fixed number of N_o obfuscators in the population who want to hide their opinions, and $N - N_o$ non-obfuscators who do not care if their opinions are disclosed or not. Initially (i.e., *stage 0*), each agent (both obfuscators and non-obfuscators)

is randomly assigned an opinion from the opinion set O, based on which she chooses an action from the action set A according to the rule of updating actions (the rule will be given in Section 4.3.1 and 4.3.2).

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

4.3.1. BEHAVIOR OF NON-OBFUSCATORS

(0) CHOOSING ACTIONS BASED ON THE INITIAL OPINIONS

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

$$P(a_g|o_k) = \begin{cases} 1 & \text{if } s_{kg} = 1 \\ 0 & \text{if } s_{kg} = -1 \\ \frac{1}{W} & \text{if } s_{kg} = 0 \end{cases} \tag{4.1}$$

(1) OBSERVING ACTIONS AND INFERRING OPINIONS

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

$$P^{(i)}(o^{(j)} = o_k|a^{(j)}) = \frac{P(a^{(j)}|o_k)}{\sum_{z=1}^{K} P(a^{(j)}|o_z)} \tag{4.2}$$

where $P(a^{(j)}|o_z)$ is the probability of choosing $a^{(j)}$ when holding opinion o_z, and can be calculated by equation (4.1). We can derive equation (4.2) from the Bayes' rule by assuming the prior probability $P(o_z) = \frac{1}{K}$ for all $z = 1, 2, \ldots, K$. The rationale behind this assumption is that agents have no prior knowledge about which opinion is more likely to be adopted by their neighbors before observing their actions.

(2) UPDATING OPINIONS

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

$$\hat{P}^{(i)}(o_k) = \frac{\sum_{j \in M_i} P^{(i)}(o^{(j)} = o_k|a^{(j)})}{\sum_{z=1}^{K} \sum_{j \in M_i} P^{(i)}(o^{(j)} = o_z|a^{(j)})}, \qquad k = 1, 2, \ldots K \tag{4.3}$$

where M_i is the set of all agent i's neighbors. As a result of positive social influence (Flache et al., 2017), agent i will update her opinion to o_k with probability $\hat{P}^{(i)}(o_k)$. In case other forms of social influence or mechanism are preferred, modelers can easily modify equation (4.3) accordingly.

(3) UPDATING ACTIONS

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

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

4.3.2. BEHAVIOR OF OBFUSCATORS

The behavior of an obfuscator is the same as a non-obfuscator in stage 1 and 2. The only difference lies in the rule of choosing actions, which applies to both stage 0 and 3. First of all, an obfuscator is still governed by the action-opinion relation: she must choose the obliged action and cannot choose the forbidden action. As a result, an obfuscator can only play obfuscation when her opinion permits more than one action. Among all the actions that are permitted by her opinion, according to Chorus et al. (2021), an obfuscator chooses the (permitted) action that reveals as little information as possible about the opinion by maximizing the uncertainty of her decision, measured by the Shannon entropy. For each action a_g, the Shannon entropy is calculated by:

$$H(a_g) = -\sum_{k=1}^{K} P(o_k|a_g)\log(P(o_k|a_g)) \tag{4.4}$$

where $P(o_k|a_g)$ is short for $P^{(i)}(o^{(j)} = o_k|a^{(j)} = a_g)$, thus it can be calculated by equation (4.2). Larger entropy implies more uncertainty. If $H(a_g) = 0$ (i.e., the entropy is minimized), choosing a_g reveals the full amount of information regarding the invisible opinion. To support this claim, we must show that $H(a_g) = 0$ only if there exists a $k = k^*$ such that $P(o_{k^*}|a_g) = 1$, and $P(o_k|a_g) = 0$ for all $k \neq k^*$. Fortunately, this has been proven by Shannon (1948) as a basic property of the Shannon entropy. Meanwhile, the entropy is maximized, according to equation (4.4), when $P(o_m|a_g) = P(o_n|a_g)$ for all $m, n = 1, \ldots, K$, that is, the observer has no knowledge about which opinion is more likely to be the opinion of an agent choosing a_g. In practice, this perfect maximization is not always achievable due to the restriction of the action-opinion matrix.

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

$$a^{(i)} = \arg\max_{a_g \in A_i} H(a_g) \tag{4.5}$$

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

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

Figure 4.1 gives a brief summary of the model. Firstly, each agent is randomly assigned an opinion. Then obfuscators and non-obfuscators choose actions based on different rules (stage 0). Afterward, a random agent is selected to update her opinion and action through a three-stage process: inferring opinions of others (stage 1), updating opinion based on the inference (stage 2), and updating action based on the updated opinion (stage 3).

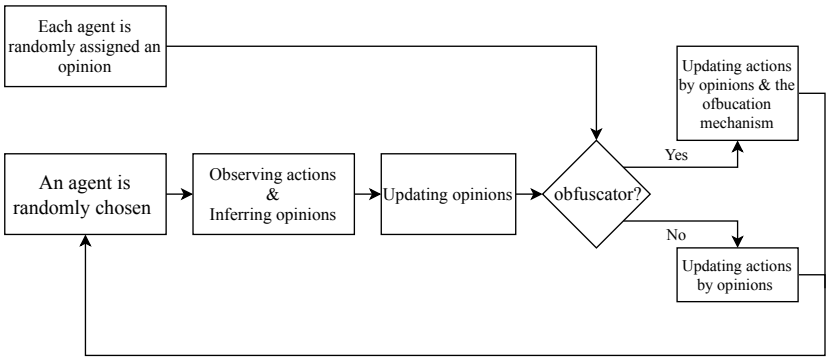


Figure 4.1: Illustration of the model.

4.4. ILLUSTRATIVE EXAMPLES

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

4.4.1. THE BATTLE BETWEEN VEGETARIANS AND OMNIVORES

We first look into a very simple case, the vegetarian-omnivore example mentioned in Section 4.1. Here we assume that there are $N = 10$ friends going to the restaurant. Ini-

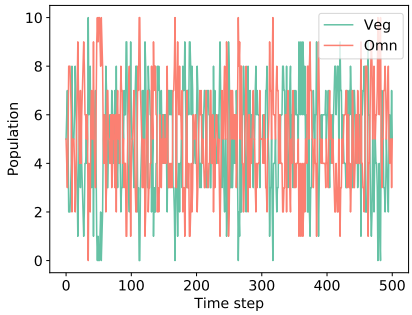
tially, $N_{Veg} = 5$ of them are vegetarians and $N_{Omn} = 5$ of them are omnivores. Given the relatively small population, it is reasonable to assume that everyone can observe the action of everyone else. We summarize this case by the following parameters and conditions: $N = 10$, G is a complete graph (i.e., everyone is a neighbor of everyone else), $A = \{a_1 = \text{(Choose) Steak}, a_2 = \text{(Choose) Salad}\}$, $O = \{o_1 = \text{Veg}, o_2 = \text{Omn}\}$ ("Veg" is short for "Vegetarian", and "Omn" is short for "Omnivore"), and the action-opinion matrix S^{VO} ("VO" stands for "the battle between **V**egetarians and **O**mnivores"):

$$S^{VO} = \begin{matrix} \\ \text{Veg} \\ \text{Omn} \end{matrix} \begin{matrix} \begin{matrix} \text{Steak} & \text{Salad} \end{matrix} \\ \begin{pmatrix} -1 & +1 \\ 0 & 0 \end{pmatrix} \end{matrix}$$

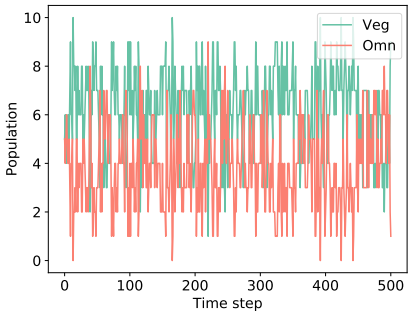
which means a vegetarian is prohibited from choosing steak and can only choose salad, while an omnivore can choose between steak and salad. Without any calculation, we can already see that an obfuscating omnivore will choose salad, and a non-obfuscating omnivore will choose randomly (i.e., flip a coin)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

The running time is set to be 500, which is sufficiently long to reach a stable outcome. Figure 4.2 shows how the number of obfuscators in the population affects public opinion, based on which we can conclude that obfuscation, in this particular case, suppresses the spread of omnivorism and promotes the popularity of vegetarianism. However, the effect is bounded: in Figure 4.2(d), even if everyone is an obfuscator, in equilibrium, there still exist a few omnivores (around 1 to 2), implying that obfuscation cannot completely eliminate the existence of omnivorism.

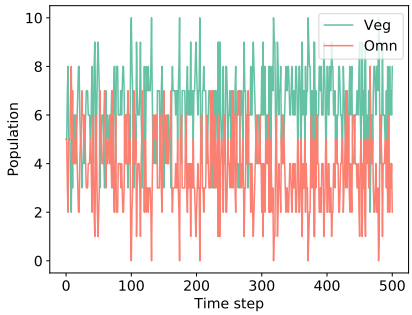
To further explore the relation between obfuscation and public opinion, we run the simulation 100 times for each N_o. In Figure 4.3, $\bar{f}_{Veg}$ (i.e., the y axis) is the fraction of vegetarians in the population averaged over the last 50 time steps of each simulation realization. Compared to Figure 4.2, Figure 4.3 provides more details. We can see that although obfuscation (represented by N_o) has a significant impact on public opinion (represented by $\bar{f}_{Veg}$), it is not a fully determining factor, as there remains a remarkable degree of variation across realizations regardless of N_o. This statement comes from the observation that even if all conditions and parameters (including N_o) are the same, the public opinion in each realization can be very different. For example, when $N_o = 0$, the lowest $\bar{f}_{Veg}$ is close to 0.35, and the highest is about 0.6. However, there is a trend that this variation decreases as the number of obfuscators increases. As another interesting finding, we can conclude that obfuscation is likely to reduce the variation in public opinion for this particular case.



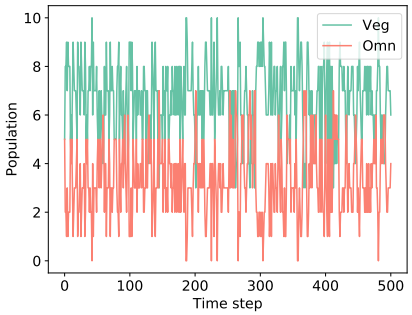
(a) $N_o = 0$



(b) $N_o = 5$



(c) $N_o = 7$



(d) $N_o = 10$

Figure 4.2: The battle between vegetarians and omnivores: population of believers in each opinion versus time step.

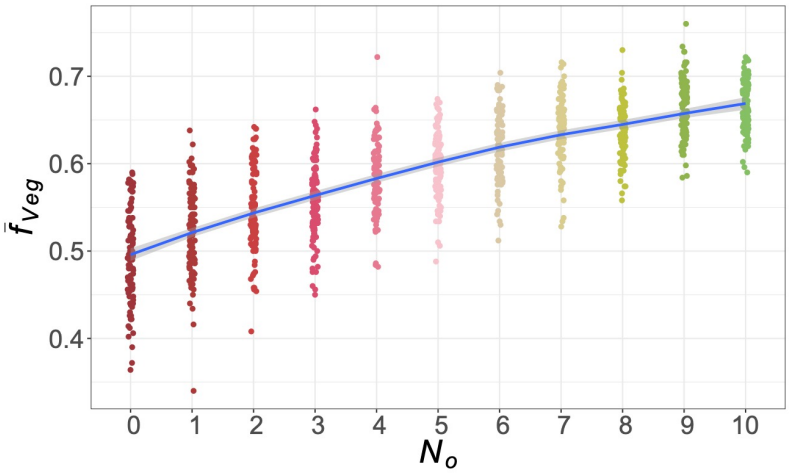


Figure 4.3: The battle between vegetarians and omnivores: fraction of vegetarians ($\bar{f}_{Veg}$) versus the number of obfuscators (N_o). For each N_o, we run 100 realizations of the simulation. $\bar{f}_{Veg}$ is obtained by averaging the fraction of vegetarians in the last 50 time steps of a realization. Each data point represents one realization. The horizontal position of each data point is slightly adjusted in order to reduce overlap. The line across the figure is the smoothed conditional mean, and the shaded area indicates the 95% confidence interval.

The rationale behind the simulation result lies in the discrepancy in observer's inference and the reality: from equation (4.2) and S^{VO}, we know that observers believe a salad-eating agent is an omnivore with a probability of 1/3. However, because an obfuscating omnivore always chooses salads, this probability is in fact larger than 1/3 in a population containing obfuscators. In the extreme case where everyone is an obfuscator, the probability increases to 1/2, the same as the probability that a salad-eating agent is a vegetarian. Such a discrepancy leads to an underestimation of the population of omnivores (or, equivalently, overestimation of the population of vegetarians). Consequently, vegetarianism becomes more popular than omnivorism because of positive social influence. At the same time, omnivorism will not go extinct because observers believe that omnivores are always likely to exist (with a relatively small probability) even if everyone chooses salads.

The rationale described above is formally expressed in the Appendix, where analytical results of this example are derived. According to the derivation, the fraction of vegetarians, averaged over all realizations (trajectories) of the dynamics, should converge to $\frac{1+\theta}{2+\theta}$ over time, where $\theta = \frac{N_o}{N}$ is the fraction of obfuscators in the population. This conclusion is validated by the simulation result in Figure 4.3, where the average $\bar{f}_{Veg}$ (averaged over 100 independent realizations) is well approximated by $\frac{1+\theta}{2+\theta}$. Furthermore, the derivation shows that the average $\bar{f}_{Veg}$ only depends on the fraction of obfuscators (θ) and the action-opinion matrix (S^{VO}), while other conditions such as the size of the population (N) and the initial distribution of each opinion (N_{Veg} and N_{Omn}) are irrelevant. Unexpectedly, the number of neighbors of each agent is also irrelevant, as long as everyone has the same number of neighbors[3].

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

$$S^{VC} = \begin{matrix} \\ \text{Veg} \\ \text{Car} \end{matrix} \begin{matrix} \begin{matrix} \text{Steak} & \text{Salad} \end{matrix} \\ \begin{pmatrix} -1 & +1 \\ +1 & -1 \end{pmatrix} \end{matrix}$$

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

4.4.2. THE EMPEROR'S NEW CLOTHES

The Emperor's New Clothes is a famous tale written by Hans Christian Andersen in 1837. The general plot is about how two swindlers pretending to be weavers, convince the

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

Emperor that the suit of clothes they made is invisible to stupid people. Everyone in the country, after observing the naked Emperor, out of fear of being considered stupid, pretends that they could see the clothes, until a child speaks out the truth.

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

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

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

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

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

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

Relevant actions, as one can imagine, include:

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

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

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

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

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

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

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

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

Observing more people choosing a_3 makes citizens with o_2 believe that there are more citizens with o_1 (i.e., $\hat{P}^{(i)}(o_1)$ increases, where i denotes citizens with o_2). However, they

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

cannot change their opinion from o_2 to o_1; therefore observing a_3 only makes them more confident in their current opinion o_2.

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

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

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

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

- Citizens with o_2: their perception of the action-opinion relation is encoded in $S^{EC}(o_2)$. Non-obfuscators choose between a_2 and a_3 with equal probability according to $S^{EC}(o_2)$; obfuscators choose a_3 according to $S^{EC}(o_1|o_2)$. For both non-obfuscators and obfuscators, observing someone choosing a_3 makes them more confident in their current opinion. The inferring process after observing other actions (i.e., stage 1) relies on $S^{EC}(o_2)$ as described in Section 4.3.

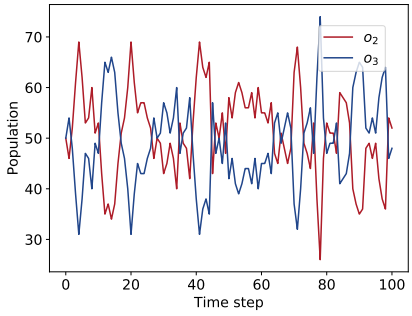
- Citizens with o_3: their perception of the action-opinion relation is encoded in $S^{EC}(o_3)$. Non-obfuscators with o_3 will choose between a_1 and a_2 with equal probability according to $S^{EC}(o_3)$; obfuscators with o_3 will choose a_2 according to $S^{EC}(o_2)$. The *observing actions and inferring opinions* process (i.e., stage 1) relies on $S^{EC}(o_3)$ as described in Section 4.3.

Readers must have realized that this case seems to be more complex than what we presented in Section 4.3. This is because the assumption that agents have the same action-opinion matrix has been relaxed. As we have discussed above, citizens with different opinions now have different perceptions of the action-opinion relation in this system. This is because citizens with o_2 have an imaginary type of neighbors: citizens with o_1. In addition, due to the different sources of social pressure (i.e., the motivation for hiding one's opinion through obfuscation), citizens also rely on different action-opinion matrices to decide how to obfuscate. Namely, obfuscators with o_2 believe that the pressure comes from citizens with o_1; therefore they choose actions according to the perception of these imaginary neighbors $S^{EC}(o_1|o_2)$. Meanwhile, obfuscators with o_3 believe that the pressure comes from citizens with o_2; therefore they rely on $S^{EC}(o_2)$ to hide opinions.

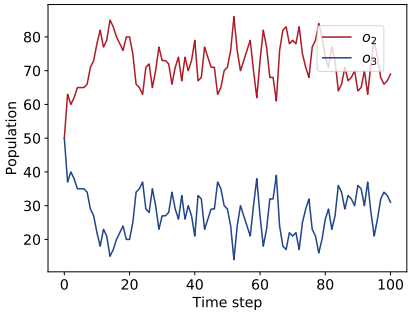
Indeed, the assumption that everyone knows and uses the same action-opinion matrix significantly simplifies the modeling process. Such a simplification is reasonable in many situations such as the vegetarian-omnivore case (Section 4.4.1), but here we show that it can be relaxed in order to capture the special mind-sets of different types of citizens.

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

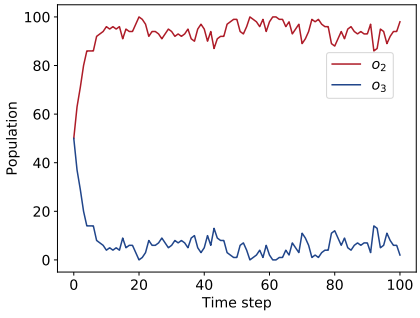
Under a set of reasonable parameters and conditions, including (1) total population $N = 100$, (2) initial population of believers in o_2 and o_3 are equal, and (3) everyone knows everyone else in the system (i.e., G is a complete graph), we obtain the simulation results shown in Figure 4.4 (dynamics of opinions) and Figure 4.5 (dynamics of actions). If none of the citizens obfuscates (i.e., $N_o = 0$), o_2 and o_3 will dominate the population in turn, and the average population believing in each opinion over time is half of the whole population (Figure 4.4(a)). Meanwhile, about half of the population will keep silent (a_2), and the rest of the population is, on average, equally divided between citizens who mock the Emperor (a_1) and admire the Emperor (a_3) (Figure 4.5(a)).



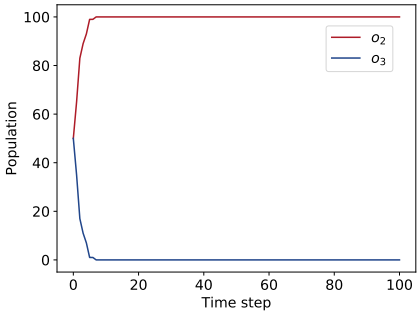
(a) $N_o = 0$



(b) $N_o = 50$

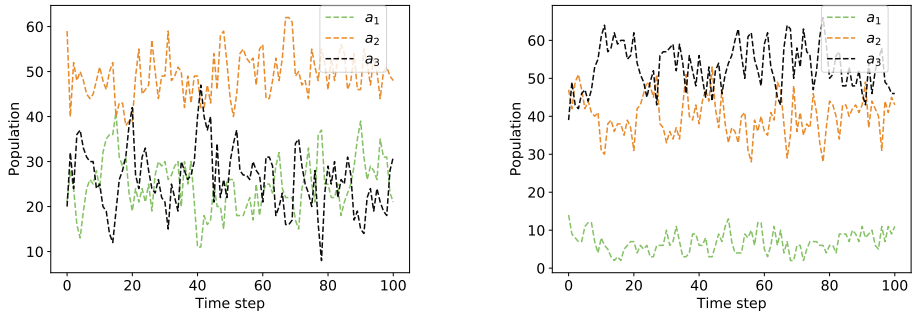


(c) $N_o = 90$

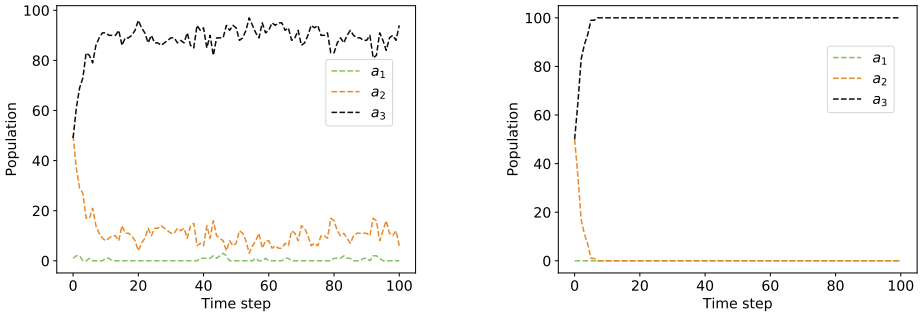


(d) $N_o = 100$

Figure 4.4: The Emperor's new clothes: opinion dynamics of the citizens. Population of believers in o_1 is always zero and thus is not plotted.



(a) $N_o = 0$ (b) $N_o = 50$



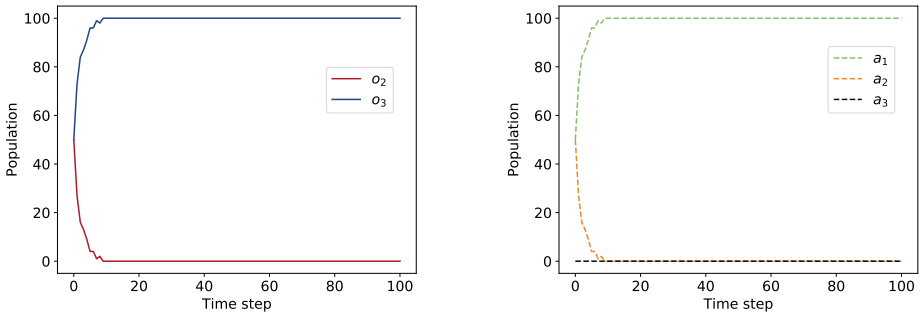
(c) $N_o = 90$ (d) $N_o = 100$

Figure 4.5: The Emperor's new clothes: population dynamics of actions chosen by the citizens versus time step.

To conclude, if no one wants to obfuscate, there is still a considerable number of citizens mocking the Emperor even when the majority is silent. However, as the number of obfuscators increases, the popularity of o_2 gradually grows (Figure 4.4(b), Figure 4.4(c)). When everyone becomes an obfuscator ($N_o = 100$), it only takes a few time steps (note that each time step contains 100 individual updates) for the whole population to reach a consensus of o_2 (Figure 4.4(d)), that is, everyone becomes the "true believer" in Centola's model, and the "public lie" becomes the "(false) public opinion" that the Emperor is dressed. In terms of actions, everyone will eventually admire the Emperor when everyone obfuscates (Figure 4.5(d)).

Now let's consider the other extreme: what if citizens, opposite to obfuscation, would like to be as transparent as possible to observers? In other words, what if citizens want their opinions to be correctly and clearly known by others? Transparent citizens with o_2 will choose a_2 according to $S^{EC}(o_2)$: although a_3 has a smaller entropy, it is misleading because it signals that the underlying opinion is more likely to be o_1. They rely on their own perception $S^{EC}(o_2)$ instead of $S^{EC}(o_1|o_2)$ (as the obfuscators do) because transparency is usually not directly related to the pressures from others. Meanwhile, transparent citizens with o_3, according to $S^{EC}(o_3)$, will choose a_1 because it directly signals that the underlying

opinion is o_3. A population full of transparent citizens, with the same parameters and initial conditions as in Figure 4.4 and Figure 4.5, would produce a completely different result (Figure 6): in a few time steps, everyone will believe that the Emperor is naked (o_3), and mock the Emperor (a_1). In the context of Centola's model, this means everyone is now a "disbeliever", and the "public lie" is replaced by the "public truth".



(a) Dynamics of opinions

(b) Dynamics of actions

Figure 4.6: The Emperor's new clothes: population dynamics of opinions and actions chosen by the citizens. Every citizen is transparent.

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

4.4.3. QUALITATIVE CONCLUSIONS FROM THE EXAMPLES

These two examples validate our early judgment that a universally correct answer to "how obfuscation affects public opinion" does not exist, but there are still some qualitative conclusions that worth mentioning. From the first example, we can arrive at a hypothesis that if an opinion only allows one action (vegetarianism in this example), it will be generally more popular than others in the presence of obfuscators. As argued in Section 4.4.1, this can be attributed to observer's overestimation of the popularity of this opinion. Although the hypothesis is not applicable in the second example (because o_1, the opinion that allows only one action, is imaginary), a similar logic can help us understand the simulation outcome. Obfuscators believing in o_2 play the same role as believers in o_1 because they always choose the same action a_3, and therefore we could conceptually divide o_2 into two categories: o_2 that is believed by obfuscators (denoted by "obfuscating o_2") and o_2 that

is believed by non-obfuscators (denoted by "non-obfuscating o_2"). Obfuscating o_2 can be viewed as an opinion that only allows one action, and hence is expected to be more popular than other opinions (such as non-obfuscating o_2) according to the hypothesis. As a result, given the total number of believers in o_2 at any instance, the more people believe in obfuscating o_2, the more popular o_2 will be in the future. Meanwhile, the number of obfuscators (N_o) determines the initial number of believers in obfuscating o_2, and therefore is positively related to the popularity of o_2.

4.5. CONCLUSION AND DISCUSSION

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

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

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

ACKNOWLEDGMENTS

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

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] Bjerring, J. C., Hansen, J. U., & Pedersen, N. J. L. L. (2014). On the rationality of pluralistic ignorance. Synthese, 191(11), 2445-2470.

[3] Blume, A., & Board, O. (2014). Intentional vagueness. Erkenntnis, 79(4), 855-899.

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

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

[6] Çelen, B., & Kariv, S. (2004). Observational learning under imperfect information. Games and Economic Behavior, 47(1), 72-86.

[7] Centola, D., Willer, R., & Macy, M. (2005). The emperor's dilemma: A computational model of self-enforcing norms. American Journal of Sociology, 110(4), 1009-1040.

[8] Chorus, C., van Cranenburgh, S., Daniel, A. M., Sandorf, E. D., Sobhani, A., & Szép, T. (2021). Obfuscation maximization-based decision-making: Theory, methodology and first empirical evidence. Mathematical Social Sciences, 109, 28-44.

[9] Davis, R. C. (2019). Obfuscating authorship: Results of a user study on nondescript, a digital privacy tool. Working Paper.

[10] De Jaegher, K. (2003). A game-theoretic rationale for vagueness. Linguistics and Philosophy, 26(5), 637-659.

[11] Dewan, T., & Myatt, D. P. (2008). The qualities of leadership: Direction, communication, and obfuscation. American Political Science Review, 102(3), 351-368.

[12] Doyle, T. (2018). Privacy, obfuscation, and propertization. IFLA journal, 44(3), 229-239.

[13] Eisenberg, E. M. (1984). Ambiguity as strategy in organizational communication. Communication Monographs, 51(3), 227-242.

[14] Fan, K., & Pedrycz, W. (2015). Emergence and spread of extremist opinions. Physica A: Statistical Mechanics and its Applications, 436, 87-97.

[15] Fan, K., & Pedrycz, W. (2016). Opinion evolution influenced by informed agents. Physica A: Statistical Mechanics and its Applications, 462, 431-441.

[16] 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).

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

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

[19] Golub, B., & Jackson, M. O. (2010). Naive learning in social networks and the wisdom of crowds. American Economic Journal: Microeconomics, 2(1), 112-49.

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

[21] 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).

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

[23] Jadbabaie, A., Molavi, P., Sandroni, A., & Tahbaz-Salehi, A. (2012). Non-Bayesian social learning. Games and Economic Behavior, 76(1), 210-225.

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

[25] Lobel, I., Acemoglu, D., Dahleh, M., & Ozdaglar, A. (2009). Rate of convergence of learning in social networks. In Proceedings of the American Control Conference (pp. 2825-2830).

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

[27] 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.

[28] Mossel, E., Olsman, N., & Tamuz, O. (2016). Efficient bayesian learning in social networks with gaussian estimators. In 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton) (pp. 425-432).

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

[30] 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.

[31] Rossman, G. (2014). Obfuscatory relational work and disreputable exchange. Sociological Theory, 32(1), 43-63.

[32] Serra-Garcia, M., Van Damme, E., & Potters, J. (2011). Hiding an inconvenient truth: Lies and vagueness. Games and Economic Behavior, 73(1), 244-261.

[33] Schilke, O., & Rossman, G. (2018). It's only wrong if it's transactional: Moral perceptions of obfuscated exchange. American Sociological Review, 83(6), 1079-1107.

[34] Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27(3), 379-423.

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

[36] 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.

[37] Takeuchi, D., Tanaka, G., Fujie, R., & Suzuki, H. (2015). Public opinion formation with the spiral of silence on complex social networks. Nonlinear Theory and Its Applications, IEICE, 6(1), 15-25.

[38] 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).

[39] Wherry, F. F., Seefeldt, K. S., & Alvarez, A. S. (2019). To lend or not to lend to friends and kin: Awkwardness, obfuscation, and negative reciprocity. Social Forces, 98(2), 753-793.

[40] You, I., & Yim, K. (2010). Malware obfuscation techniques: A brief survey. In 2010 International Conference on Broadband, Wireless Computing, Communication and Applications (pp.297-300).

APPENDIX

In this Appendix, we derive an analytical solution of the dynamics described in Section 4.4.1, namely the Vegetarian-Omnivore example, whose simulation results have already been given in Figure 4.2 and Figure 4.3. The derivation closely follows Tang and Chorus (2019)[7], which itself is an extension of a typical derivation of the voter model (Krapivsky et al., 2010). It should be noted that the derivation significantly benefits from the simplicity of S^{VO}, and is therefore only applicable to this particular case.

Recall the action-opinion matrix used in the example:

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

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

The probability that agent i changes her opinion ("flip rate"), based on equation (4.2) and (4.3), can be written as:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Note that $R(i) \equiv < o^{(i)} > = \sum_o P(o^{(i)} = o) o$ and o can only take two values ± 1; we have:

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

Similarly, we have $R^*(j) = < P(o^{(j)} = 1 | a^{(j)}) - P(o^{(j)} = -1 | a^{(j)}) >$, hence:

$$R^*(j) = 2 < P(o^{(j)} = 1 | a^{(j)}) > -1 \tag{A8}$$

According to the definition of "$< \cdot >$":

$$< P(o^{(j)} = 1 | a^{(j)}) > = P(o^{(j)} = 1 | a^{(j)} = a_1) P(a^{(j)} = a_1) + P(o^{(j)} = 1 | a^{(j)} = a_2) P(a^{(j)} = a_2) \tag{A9}$$

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

$$< P(o^{(j)} = 1 | a^{(j)}) > = \frac{2}{3} P(a^{(j)} = a_2) \tag{A10}$$

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

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

Note that:

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

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

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

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

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

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

and

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

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

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

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

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

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

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

From (A7) we know $R(i) = 2P(o^{(i)} = 1) - 1$. Therefore:

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

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

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

The stable fixed point of (A20) is:

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

which is equivalent to:

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

(A22) tells us that at equilibrium, the share of vegetarians in the population is $\frac{1+\theta}{2+\theta}$. However, this does not mean for every realization, the system will converge to this equilibrium. Instead, the share of vegetarians averaged over all realizations of the dynamics will converge to $\frac{1+\theta}{2+\theta}$. This result is in line with Figure 4.3, where $\bar{f}_{Veg}$ averaged over all the realizations carried out in the simulation (which is not "all realizations" but an approximation of "all realizations") is well approximated by $\frac{1+\theta}{2+\theta}$.

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.

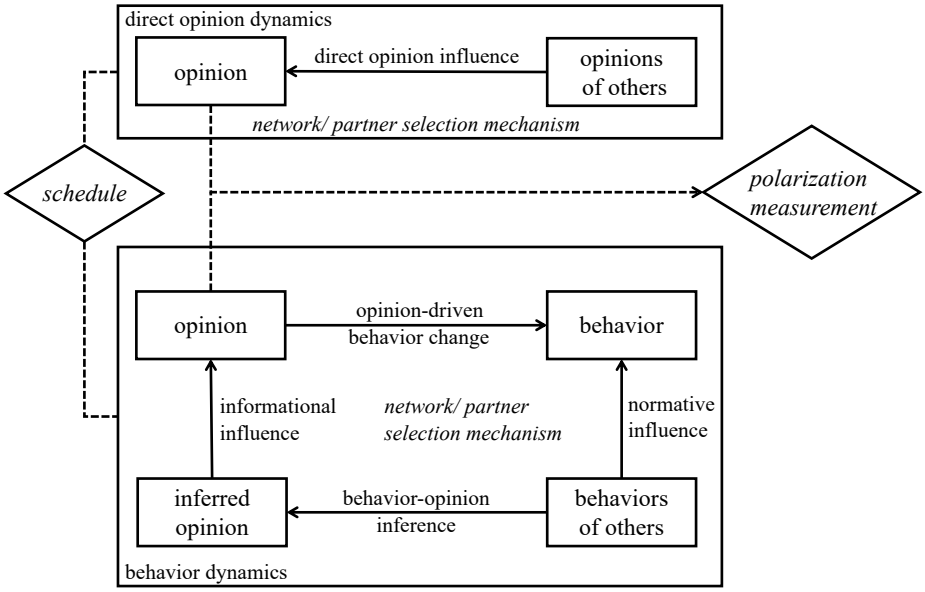


Figure 5.1: Structure of the MOBEP framework.

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:

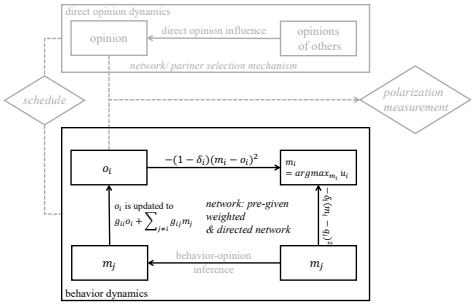
- **Opinion:** a discrete variable o_k. An agent i may take one of the possible opinions at a time: $o^{(i)} \in \{o_k\}_{k=1,...,K}$.
- **Behavior:** a discrete variable m_g (called "action"). An agent i may take one of the possible actions at a time: $m^{(i)} \in \{m_g\}_{g=1,...,G}$.
- **Opinions 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.

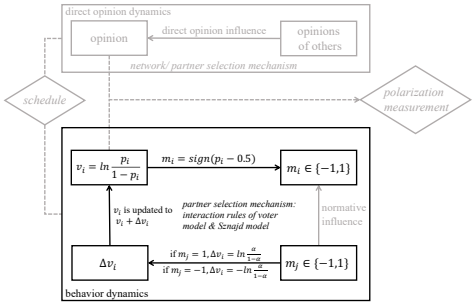
SNOAEs: Zhan et al. (2021)'s SNOAEs model can be decomposed as follows according to the framework:

- **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.

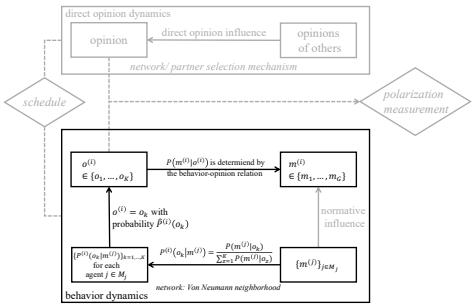
The fittings of the models to the framework are presented in Figure 5.2, which also shows one of the functionalities of the framework – organizing literature (see Section 5.3.5):



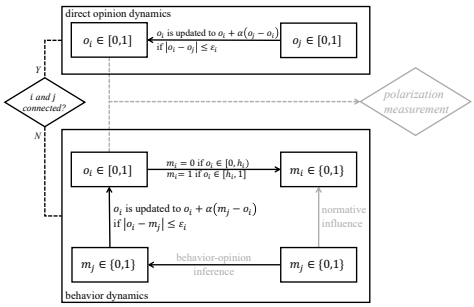
(a) Buechel et al.'s model of conformity



(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.

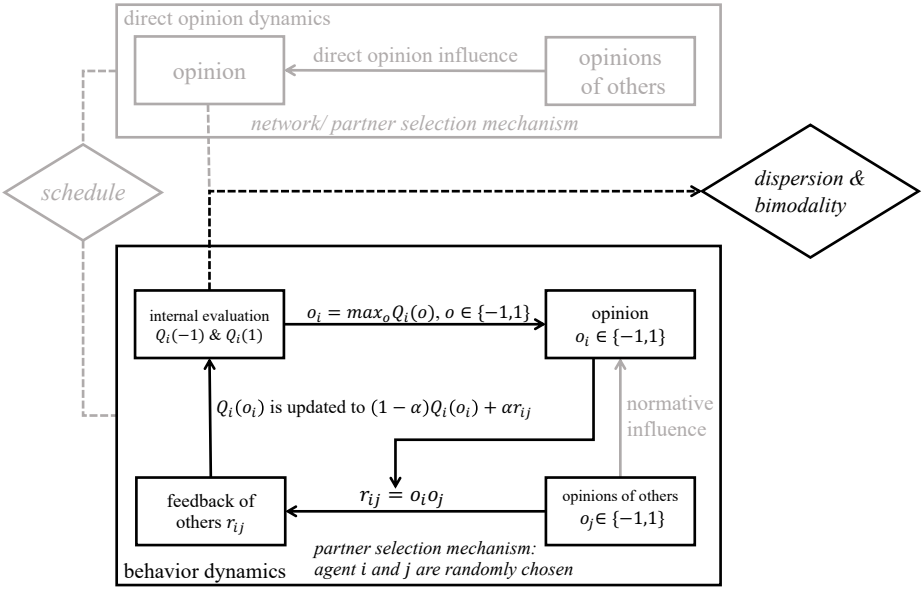


Figure 5.3: Fitting the social feedback model to the MOBEP framework. Framework components that do not appear in the model are marked gray. Additional information: $o \in \{-1,1\}$, and o_i can take one of the values of o. α is the learning rate.

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).

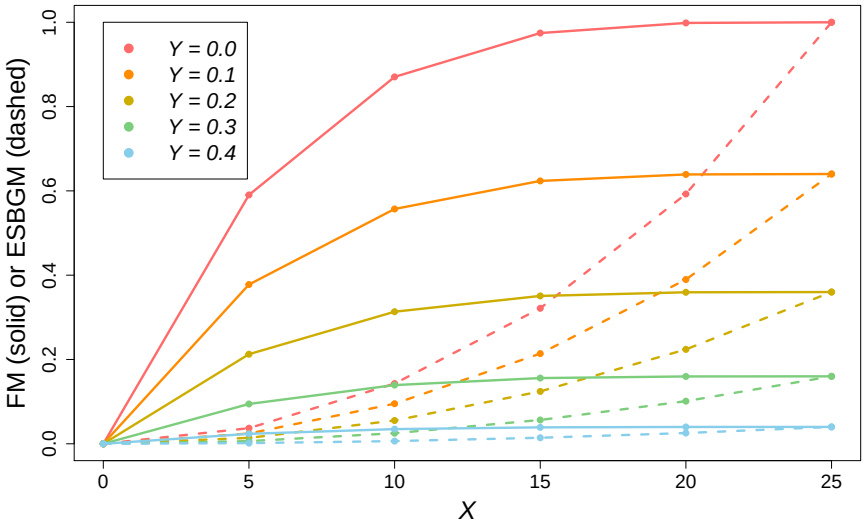


Figure 5.4: Comparison between FM (solid) and ESBGM (dashed).

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.

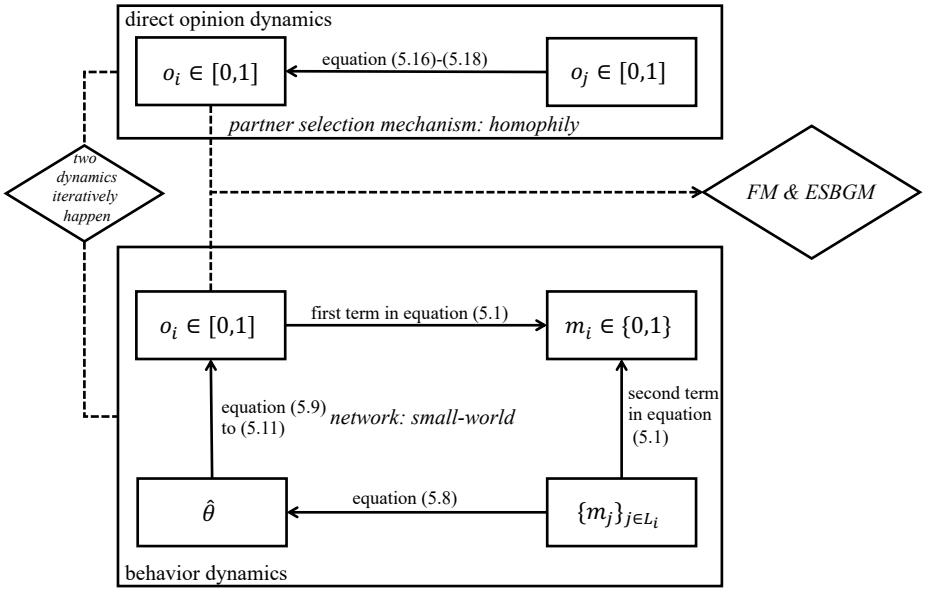


Figure 5.5: Fitting the case study model to the MOBEP framework.

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 $\lambda = 0.5$.

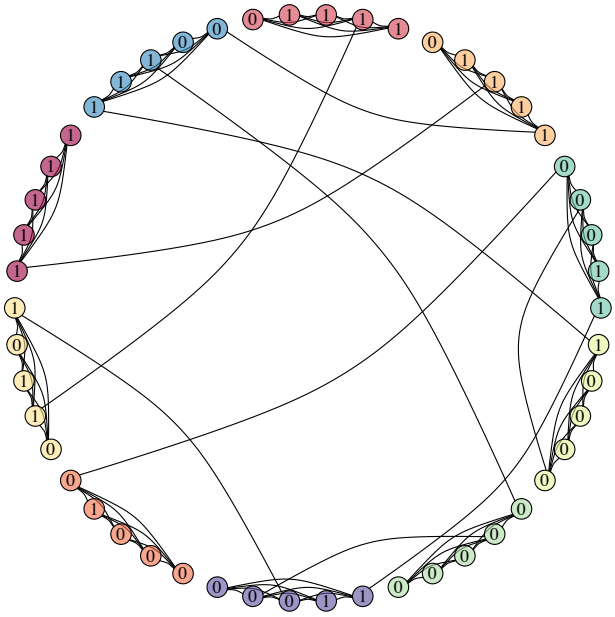


Figure 5.6: Opinion distribution on a small-world network at $t = 1000$ of a typical run with only behavior dynamics. A circle represents an agent, where the number inside is the agent's opinion, rounded to the second decimal place (the rounded results are either 0.00 or 1.00 so the parts after the decimal 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.

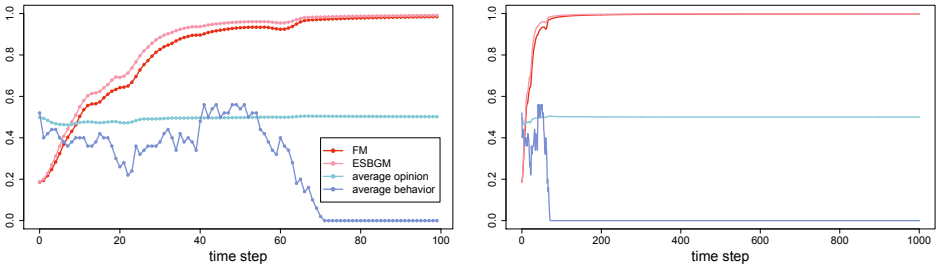
The dynamics of this particular run is described by the evolution of FM, ESBGM, average opinion, and average behavior (Figure 5.7). While average opinion remains relatively stable, opinion polarization levels (FM and ESBGM) rise dramatically in the first 100 time steps, and average behavior reaches zero (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.

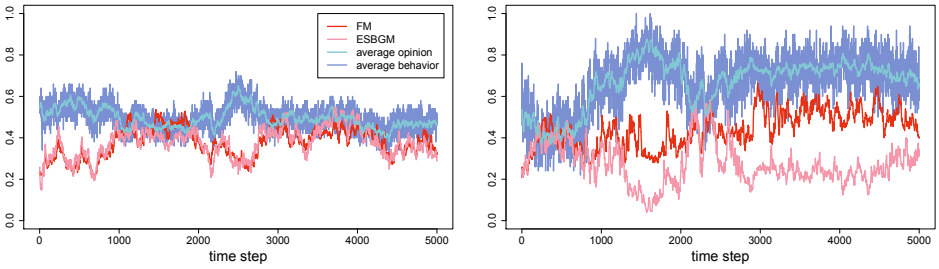
Very roughly, one time step in simulation is of a similar scale of one day in real life. Therefore, choosing an extremely large T (say, $T > 5000$) would be unrealistic considering that many relevant factors that are excluded from the model (such as the infection rate, development of vaccines, and mask 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

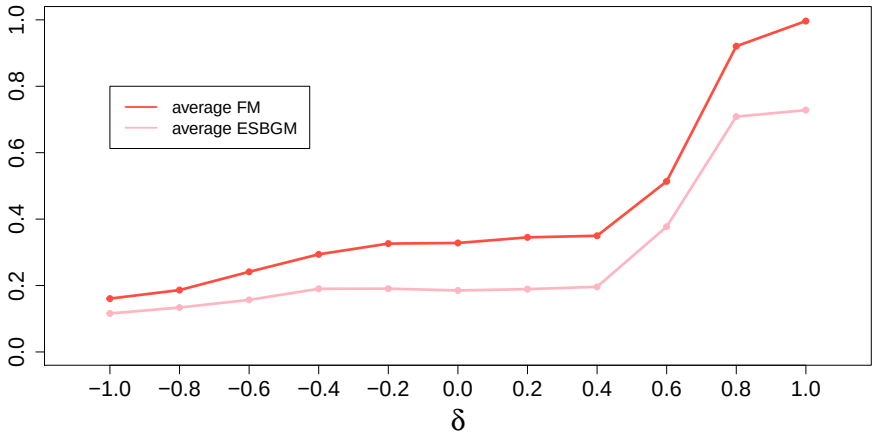


Figure 5.9: Polarization level (FM & ESBG) vs conformity preference (δ) in behavior dynamics. Average FM and ESBGM are obtained from the system at $t = 1000, 2000, 3000, 4000$, and 5000. Parameters include: $N = 50$, $s = 5$, $p = 0.2$, $\delta_i = \delta \ \forall i$, and $\lambda = 0.5$.

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.

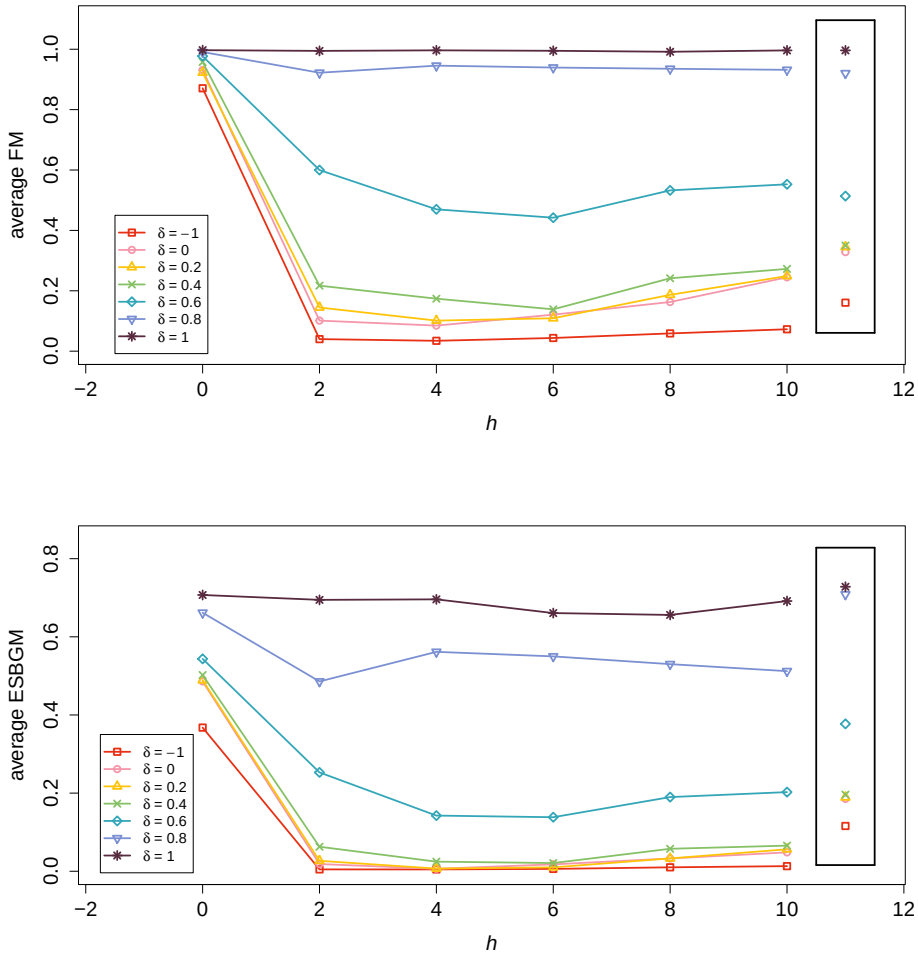


Figure 5.10: Effect of homophily h on opinion polarization measured by FM and ESBGM under different values of universal conformity preference δ averaged between $t = 1000, 2000, 3000, 4000$, and 5000. Points in the box on the right-hand side are the polarization levels in behavior dynamics only (i.e., 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).

6

CONCLUSIONS AND REFLECTIONS

Opinion polarization is a widespread and potentially dangerous social phenomenon that is now reshaping economies and societies around the world. This statement, which is now part of common sense, has been repeatedly endorsed by real-life events of multiple scales. The current riots between pro- and anti-vaccine camps provide the most recent worldwide endorsement, underlining the urgency of countering opinion polarization, especially in crises such as the COVID-19 pandemic.

Battling against opinion polarization requires scholarly efforts to help us understand what polarization is and what causes polarization. Echoing this requirement, the thesis aims to make methodological contributions to opinion polarization studies from a rarely explored perspective, which centers on the "private and unobservable" feature of opinions. The main idea is that opinions cannot be directly observed by others but need opinion-driven behaviors to mediate their interactions with the opinions of others. In order to help modelers put this idea into scientific practice, the thesis proposes four research methods that cover different parts of opinion polarization, with the ultimate goal of arriving at a more specific and realistic understanding of this mysterious phenomenon.

This chapter concludes the thesis by summarizing and discussing these methods. Section 6.1 will provide an overview of each method while highlighting their functionalities and positions in the literature. Section 6.2 will unveil the relations between the four methods, and present them as an integrated toolkit. Avenues for future studies will be discussed in Section 6.3, followed by the societal relevance in Section 6.4. Finally, Section 6.5 reflects on various issues that emerge from the research process, including my thoughts on the field of opinion dynamics/ polarization.

6.1. OVERVIEW OF THE METHODS PRESENTED IN THE THESIS

In Chapter 1, I have declared that the primary research goal of the thesis is "to develop methods that could systematically support the exploration of mechanisms that can explain the polarization of private and unobservable opinions", which can be further divided into four sub-goals. From Chapter 2 to 5, each chapter aims to achieve one of the sub-

goals by presenting a research method for opinion polarization studies. The relations between the sub-goals, chapters, and methods are summarized as follows:

- Sub-goal 1: To develop a formal and broadly applicable polarization measurement that is coherent with the notion of group.
  - corresponding chapter: Chapter 2.
  - corresponding method: Equal Size Binary Grouping Measurement (ESBGM).
- Sub-goal 2: To develop a modeling method that can incorporate the "private and unobservable" feature of opinions in opinion polarization mechanisms.
  - corresponding chapter: Chapter 3.
  - corresponding method: Action-Opinion Inference Model (AOI model).
- Sub-goal 3: To develop a modeling method to study the effect of the deviation from the "default strategy", such as obfuscation, on opinion polarization.
  - corresponding chapter: Chapter 4.
  - corresponding method: Obfuscation-based opinion dynamics model (obfuscation model).
- Sub-goal 4: To develop a unifying framework of the co-evolution of opinion and behavior to organize existing efforts and facilitate future works of opinion polarization mechanisms.
  - corresponding chapter: Chapter 5.
  - corresponding method: Framework of Modeling Opinion-Behavior Co-evolution for Explaining Opinion Polarization (MOBEP framework).

I opened the entire thesis by promising to fulfill these sub-goals (and by doing so, I also consequently fulfill the primary research goal) in Chapter 1, and now I can claim that the promise has been kept. In the rest of this section, I will briefly revisit these methods, together with their functionalities and positions in the literature.

6.1.1. ESBGM: A GROUP-BASED POLARIZATION MEASUREMENT

Chapter 2 presents a group-based polarization measurement derived from my newly proposed, clustering-inspired grouping method called Equal Size Binary Grouping (ESBG). The measurement is not named in Chapter 2 but is later called Equal Size Binary Grouping Measurement (ESBGM) in Chapter 5. ESBG divides the data set to be measured into two equal-size groups based on the similarities between data points, and ESBGM is built upon this group structure as a function that increases with between-group heterogeneity and decreases with within-group heterogeneity. By providing an elegant and practical manner to identify and utilize group structures, ESBGM solves the longstanding problem in polarization measurements that a proper group structure is absent. Apart from its relations with between- and within-group heterogeneity, ESBGM possesses a number of

other ideal properties, including continuity (ESBGM is a continuous function), dimensionality (ESBGM is able to deal with both uni- and multi-dimensional data), maximum & minimum (ESBGM is maximized when data points in the same group are identical, and the difference between the two groups is maximized; ESBGM is minimized when all data points are identical), and normalization (the value of ESBGM is in the range between 0 and 1).

In Chapter 2, ESBGM is compared with measurements of bimodality (e.g., Ashman's D and bimodal separation) and bipolarization (e.g., Wolfson's index). All these measurements rely on the bi-division of the data set, but they are not measuring the same thing. First, bimodality may be a necessary but not sufficient condition for polarization. Therefore, bimodality measurements only capture one of the many aspects of polarization measured by ESBGM. Second, ESBGM and bipolarization measurements represent two different lines of polarization measurement, namely the "Esteban & Ray's line" (which focuses on clusters) and "Wolfson's line" (which focuses on the disappearance of the middle class). In Chapter 5, I further compare ESBGM with the FM index – a representative polarization measurement that has been widely used in opinion dynamics studies. I find that ESBGM is more sensitive to the distribution of group sizes than the FM index, while the FM index is more sensitive to extreme opinions.

Each polarization measurement reflects a unique understanding of polarization, and I cannot judge if a measurement "universally" outperforms another. Therefore, instead of claiming ESBGM to be the best measurement, I argue that it uniquely captures the widely accepted notion that "group" is a crucial ingredient of polarization. In practice, I would suggest using multiple measurements in opinion polarization studies to obtain a broader view of the system of interest (see the case study in Chapter 5 for an example).

6.1.2. AOI MODEL: AN AGENT-BASED MODELING METHOD FOR CO-EVOLUTION OF BEHAVIOR AND OPINION

The second sub-goal asks for "a modeling method that can incorporate the private and unobservable feature of opinions in opinion polarization mechanisms". In other words, the modeling method should be able to "incorporate behaviors in opinion polarization mechanisms as a messenger of opinions". This method is then formally introduced in Chapter 3 in the form of an agent-based model called Action-Opinion Inference (AOI) model.

The AOI model is not the first opinion dynamics model to include both opinion and behavior in the same dynamics. In Chapter 3, it is compared with the CODA model. The notable difference is that opinions are continuous in the CODA model but are discrete in the AOI model. The discreteness in the AOI model makes it possible to use the deontic logic (i.e., permission, obligation, and prohibition) to describe various behavior-opinion relations, while in the CODA model, the relations between continuous behavior and discrete opinion are rather straightforward, and in some cases, oversimplified (see Section 5.2.2 of Chapter 5 for details). In Chapter 5, the MOBEP framework helps us systematically compare the AOI model with other behavior-opinion co-evolution models, especially the conformity model (Buechel et al., 2015) and the SNOAEs model (Zhan et al., 2021). Among all these models, the AOI model, as the name indicates, almost exclusively describes the behavior-opinion inference process (meaning "inferring one's opinion from her behav-

ior") in a Bayesian manner, whilst other models would simply ignore the complexity of the process by treating behaviors as "extreme" opinions[1].

6.1.3. OBFUSCATION-BASED OPINION DYNAMICS MODEL: AN AGENT-BASED MODELING METHOD FOR OBFUSCATION-BASED OPINION DYNAMICS

Obfuscation brings additional uncertainty to the dynamics of opinions, and thereby has a potentially significant role in opinion polarization. On the basis of the AOI model, Chapter 4 presents a modeling method for studying obfuscation in the form of an agent-based model. This method models obfuscation as an agent's strategy to choose the behavior that gives out the least information about her opinion while ensuring that the chosen behavior is allowed. At the other extreme, the method is also capable of modeling transparent agents who want their opinions to be accurately known by others.

Obfuscation represents strategies of "intelligent agents" that deviate from the "default strategy" of "naive agents", which is exactly what the third sub-goal asks us to model. The agents described in the AOI model are considered naive as their behaviors are purely and passively determined by their opinions, while the agents modeled by this method have a higher level of intelligence in the sense that their behaviors are also directed by their individual goals (e.g., hiding their opinions or disclosing their opinions) and knowledge about other agents (e.g., how other agents would interpret their behaviors). As stated in Chapter 4, similar concepts include deception, strategic ambiguity, and intentional vagueness, whose roles in opinion polarization are also worth exploring. In light of this argument, I can frame the obfuscation model as an "intelligent" extension of the AOI model.

Obfuscation is also a "special characteristic" (Dong et al., 2018) of agents. Other special characteristics that have been extensively studied in opinion dynamics studies include skepticism (Tsang & Larson, 2014), zealotry (Waagen et al., 2015; Verma et al., 2014; Mobilia, 2003), and stubbornness (Yildiz et al., 2013). Obfuscation differs from them as it is exclusive to behavior-included opinion dynamics (i.e., opinion-behavior co-evolution): in the absence of behavior, there is no room for agents to obfuscate. From this perspective, this method can be viewed as a novel approach to studying special characteristics of agents in opinion dynamics.

6.1.4. MOBEP FRAMEWORK: A UNIFYING FRAMEWORK OF MODELING OPINION-BEHAVIOR CO-EVOLUTION FOR EXPLAINING OPINION POLARIZATION

Chapter 5 presents an opinion-behavior co-evolution framework MOBEP (**M**odeling **O**pinion-**B**ehavior co-evolution for **E**xplaining Opinion **P**olarization), which is particularly designed to explain opinion polarization via agent-based modeling. The framework is constructed by five key components that together capture all aspects of opinion-behavior co-evolution. The functionality of the framework is twofold. First, it provides an overarching architecture to organize existing efforts devoted to modeling opinion-behavior

[1] This is possible when an agent's opinion is set to be a continuous variable ranging from 0 to 1, and her behavior is set to be a binary variable that is either 0 or 1. As a result, behaviors can be directly added to and subtracted from opinions. See Section 5.2 of Chapter 5 for a review.

co-evolution, which would advance the accumulation of insights and modeling techniques. Second, future works of opinion polarization would benefit from the framework as these insights and techniques shall provide practical guidance for building new models, and the overarching architecture will help modelers better place their works in the vast literature of opinion polarization.

In the framework, classic models of opinion dynamics that exclude behaviors are represented by one of the key components – direct opinion influence. In this sense, my framework is of a higher level of granularity than frameworks of classic opinion dynamics such as the framework proposed by Coates et al. (2018), or, in other words, the classic ones are the subsets of mine. Meanwhile, the framework is specific to opinion dynamics, and is thereby of a lower granularity level than general frameworks of agent-based models (e.g., Masad & Kazil, 2015).

6.2. RELATIONS BETWEEN THE METHODS

An intrinsic feature of this paper-based thesis is that each content chapter is relatively independent, and all the methods presented in these chapters can stand on their own feet. However, the thesis is not a collection of poems where the only connection between each chapter is the same authorship. In fact, there is an underlying thread connecting each method to form one integrated toolkit to fulfill the primary research goal.

The methods presented in Chapter 3-5 are all built upon one central notion: behaviors are the messenger of opinions (see Section 5.2.3 of Chapter 5 for details)[2]. The AOI model lays down the foundation for modeling opinion dynamics under this notion, and the obfuscation model extends the boundaries by taking into consideration agents' "deviate" strategy of obfuscation. The MOBEP framework takes one step back from specific modeling tasks (e.g., the AOI model and the obfuscation model) to obtain a bird's eye view of the opinion-behavior co-evolution. Finally, the measurement ESBGM connects these three methods to opinion polarization by measuring the results generated by them, and forms a complete toolkit that is particularly designed for opinion polarization studies.

To see how these methods work as one integrated toolkit, one first needs to understand the research approach of a typical opinion polarization study in computational sociology. Usually, such a study starts with a core assumption that is potentially important for the emergence of opinion polarization. The core assumption is then implemented in an agent-based model, together with other model components. Running simulations of the model leads to simulation results of opinion distributions, from which conclusions about the assumption will be drawn. The four methods facilitate the research approach in the following ways:

- **Creating assumptions**: Considering the potentially crucial role of behavior in opinion polarization, it is promising to use the AOI and the obfuscation model as starting points to develop core assumptions, such as (if) obfuscation can foster polarization. At the same time, the framework could inspire the creation of both core assumptions and other model components with its accumulation of insights gained from previous works.

[2] This notion is a refined version of the central notion of the AOI model ("learning opinions by observing actions"). See Chapter 3.

- **Translating assumptions to agent-based model**: The AOI model and obfuscation model can be used to translate core assumptions that are related to both opinions and behaviors into agent-based model components. The framework provides an architecture to organize all model components (including the core assumption) by helping modelers identify the relations between them. In addition, such an architecture makes it easier to modify and then employ existing implementations of these model components from previous works: for example, the case study provided in Chapter 5 benefits from different models encompassed by the framework: the opinion-driven behavior change component is modified from the conformity model (Buechel et al., 2015), and the direct opinion influence part is modified from the rejection model (Flache & Macy, 2011).

- **Translating results to conclusions**: Translating the agent-based model to simulation results requires programming skills, while translating the simulation results to conclusions requires a proper polarization measurement. A polarization measurement not only summarizes the state of the whole system but also reflects our understanding of polarization. The measurement ESBGM presented in Chapter 2 echoes the widely agreed understanding that the notion of group, instead of individual, is the crucial ingredient of polarization, and hence is an ideal tool to derive conclusions from simulation results.

To conclude, the four methods are united as one toolkit that covers almost all the steps that are required to explain opinion polarization by agent-based modeling.

Another interesting perspective to understand the relations between these methods is to consider the framework as a configuration that accommodates the other three methods. The AOI model and the obfuscation model provide valuable insights for the opinion-driven behavior change component and the behavior-opinion inference component of the framework, and naturally, ESBGM relates to the component of polarization measurement (see Section 5.3 of Chapter 5 for the framework components). From this perspective, I can alternatively conclude that the entire thesis is about the MOBEP framework presented in Chapter 5, while Chapter 2-4 are preparing relevant ingredients.

6.3. AVENUES FOR FUTURE RESEARCH

As far as I can imagine, future research based on this thesis contains at least three possible categories: (i) improving the methods proposed in the thesis, and (ii) utilizing these methods for opinion polarization studies motivated by new research questions, and (iii) developing new research methods based on the proposed methods. Despite their different relations with the thesis, the three categories shall work towards the same goal: to deepen our understanding of the polarization of "private and unobservable opinions".

6.3.1. IMPROVING THE PROPOSED METHODS

As the output of my Ph.D. research, this thesis inevitably suffers from limitations of time and is thereby far from perfection. Consequently, there are many interesting and promising avenues for follow-up research that aims to improve the methods proposed in the thesis from different aspects. Basically, such avenues may take approaches such as

testing the methods with relaxed or modified assumptions, calibrating or validating the methods according to empirical data, generalizing the methods for broader applicability, and implementing the method in a more user-friendly manner. At the end of each relevant chapter, I have already discussed a number of avenues for future research, but more have been identified after the papers were published. For the reader's convenience, they are now summarized and briefly explained as follows:

AVENUES FOR IMPROVING ESBGM

(1) Systematically comparing ESBGM with existing polarization measurements, especially in the context of various attribute distributions: In Chapter 2, I have already compared ESBGM with bimodality and bipolarization measurements, and in Chapter 5, ESBGM is compared with the FM index. In addition, there are many other widely used polarization measurements, such as the Esteban-Ray index and the SSG index, whose relations with ESBGM remain unknown. Comparing ESBGM with these measurements would help us discover its uniqueness in the vast literature, and find out in which situations ESBGM would be more suitable than its competitors. To make meaningful comparisons, it is recommended that apart from theoretical analysis (e.g., checking whether some properties, such as increased spread, are satisfied. See Section 2.6 of Chapter 2 for an example), these measurements (including ESBGM) should be applied to various types of attribute distributions in order to compare each measurement's "characteristics", namely the particular aspect(s) of a distribution that the measurement values the most in assessing opinion polarization. For example, in Chapter 5, I find out that compared to the FM index, ESBGM places more weight on extreme values than group sizes. Distributions of two or more peaks are favorable candidates as they can be used to test how each measurement responds to intermediates state between consensus and polarization.

(2) Re-examining selected polarization studies by re-measuring their results using ESBGM: The choice of measurements reflects the authors' understanding of opinion polarization, and is able to substantially affect the final results. Section 5.4.4 of Chapter 5 has provided a vivid example where the polarization levels reported by ESBGM and the FM index can be extremely different when the opinion difference is large but the group size difference is small (see Figure 5.4 of Chapter 5). Therefore, it becomes pivotal to re-examine important conclusions derived in previous polarization studies by re-measuring corresponding results using ESBGM. Current conclusions may be rewritten and new insights can be developed if there exist significant differences between the ESBGM-measured results and the existing ones, implying that a widely accepted measurement (which represents a widely accepted understanding of polarization) is essential to reach consensus about polarization mechanisms. Meanwhile, the difference could also tell if ESBGM is suitable for the context of that particular study, and if necessary, one can improve ESBGM accordingly.

(3) Implementing ESBGM based on clustering techniques other than K-means: To implement ESBGM, one first needs to divide the data set into two groups of equal sizes (i.e., implementing ESBG). In Chapter 2, I have provided a demonstration where ESBG (and hence ESBGM) is implemented based on the most famous and simple clustering technique: K-means. According to Jain (2010) and HajKacem et al. (2019), thousands

of clustering techniques have been published in the last sixty-some years, and K-means is by no means the only one of them that can be adapted for ESBGM implementation. Each clustering method has its own advantage in discovering hidden grouping structures from the data, and ESBGM can exploit these advantages by employing various clustering methods as its basis. Here I present a list of research directions that are worth noticing in this regard:

(a) **Fuzzy-clustering-based ESBGM:** A large number of clustering techniques, including K-means, are the so-called "crisp clustering", meaning that each data point exclusively belongs to one of the groups (Xu & Wunsch, 2009). This "crisp" feature is sometimes harmful to the applicability of ESBGM because exclusive group membership means it will be impossible to divide a data set into two equal-size groups if its size is an odd number. As a result of this, it remains undecided how to deal with this type of data sets ("*For the sake of simplicity, we only discuss systems whose size is an even number*", Section 2.3.2 of Chapter 2). This problem can be solved by replacing crisp clustering (e.g., K-means) with fuzzy clustering, where each data point belongs to all (in our context, two) groups with a certain probability. In other words, each data point, instead of being assigned to one of the groups, now has a membership of a certain degree (Xu & Wunsch, 2009). Such a "soft assignment" (Jain, 2010) is able to address the issue of applicability because a data set of an odd-numbered size can now be divided into equal "shares" of data points by a fuzzy clustering technique. Moreover, fuzzy clustering can eliminate group membership transitions caused by imperceptible changes of a data point near the group boundary, which is at least an intellectually unfavorable circumstance for some scholars. Introduction to fuzzy clustering can be found in most textbooks on clustering or machine learning (e.g., Xu & Wunsch, 2009; Aggarwal & Reddy, 2014; Bishop, 2006; Müller & Guido, 2016), while a number of software tools are available to perform the task in various programming languages (e.g., *fclust* [3] for R, and *fuzzy-c-means* [4] for python). Therefore, implementing ESBGM based on fuzzy clustering should be technically feasible.

(b) **Outlier-robust-clustering-based ESBGM:** The last thing we want to encounter while using ESBGM is an outlier, which may dramatically change the division of groups and thereby erode our trust in the results. Even if the outlier is not caused by mistakes, its disproportional influence on the entire system's polarization level (measured by ESBGM) is rarely favored. The problem comes from K-means' strong sensitivity to outliers (Chawla & Gionis, 2013; Olukanmi & Twala, 2017), and the solution is, straightforwardly, to employ outlier-robust clustering techniques. Several modifications of K-means have been proposed to address the problem. K-medoids undermines the influence of outliers on clustering by enforcing each centroid to be fully overlapped with one of the objects within the group (Kaufman & Rousseeuw, 1990; Bishop, 2006). K-means-- (pronounced "k means minus minus") (Chawla & Gionis, 2013) and K-means # (pronounced "k means sharp") (Olukanmi & Twala, 2017) are able to detect outliers while preserving the advantages of K-

[3] https://cran.r-project.org/web/packages/fclust/index.html
[4] https://pypi.org/project/fuzzy-c-means/

means, allowing us to decide later if or how these outliers will be included in the future measuring process.

(c) **Large-scale-clustering-based ESBGM:** To save us from unnecessary complexity, the demonstration data sets used in Chapter 2 are of relatively small scales (less than 300 objects and two dimensions), and the measuring process is extremely fast. With the development of data science, scholars now need to deal with extra large-scale data that could easily take the original K-means a significant amount of time[5] to deal with, indicating that the K-means-based ESBGM would be even slower. To speed up the measuring process, one simply needs to adopt one of the improved versions of K-means (e.g., Pelleg & Moore, 1999; Stoffel & Belkoniene, 1999; Kanungo et al., 2002), or specific clustering techniques such as CURE (Guha, 1998)[6]. The cost of improved computational efficiency is the increased complexity of implementation, which can be minimized if the relevant algorithm is presented as a "black box", such as an R package or Python module, to scholars.

AVENUES FOR IMPROVING THE AOI MODEL

(1) Combining the model with homophily: The partner selection mechanism of the original AOI model in Chapter 3 is based on the voter model, which states that the focal agent would randomly choose a neighbor to interact with. As stated in Chapter 5, compared to random selection, homophily, the tendency that like-minded people are more likely to interact with each other, is a more realistic assumption of the partner selection mechanism, and in many models plays a crucial role in determining the dynamics of the system (e.g., Mäs & Flache, 2013). Combining the AOI model with homophily should lead to significant improvements in its ability to model real-life situations.

(2) Testing the AOI model in various network structures other than Von Neumann neighborhood: The original AOI model runs on the Von Neumann neighborhood for two reasons: (i) to minimize the complexity caused by networks, and (ii) to maximize the clarity of visualization. Future studies can extend the applicability of the model by testing it in other networks such as small-world networks (which mimic real-life connections between acquaintances) and multiplex networks (for situations such as co-evolution of online and offline opinion dynamics).

(3) Analytically solving the model of general action-opinion relations: In Chapter 3, the model has been analytically solved under the simplest action-opinion relation. Despite the possible difficulty, analytically solving the model of more general relations could validate the simulation results and help us better understand the dynamics.

(4) Empirically validating the model at both micro and macro levels: Like most opinion dynamics models, the AOI model has not been empirically validated, which means its model setups (micro-level) and aggregated outcomes (macro-level) have not been compared with real data. In the field of opinion dynamics, it is a common practice to build a

[5] The time complexity of K-means is $O(Nkd)$, where N is the data size, and d is the number of data dimensions (Xu & Wunsch, 2009).

[6] At this moment I don't know if these techniques can be modified according to the requirements of ESBGM.

model purely upon theories and assumptions, but recently, a trend of empirical validation has emerged, aiming to improve the real-life reliability of the in-silico outcomes (see Section 6.5.4 for a detailed discussion). The trend implies that empirical validation should be a necessary step to improve and promote the AOI model if I would like it to be used by more scholars.

(5) Generalizing the model by adopting continuous numerical values to represent action-opinion relations: The AOI model uses matrices with deontic logic to describe action-opinion relations. An entry of the matrix, which represents the evaluation of an action by an opinion, must be either + (obligation), − (prohibition), or 0 (permission). This setup remarkably simplifies the model by dissolving the opinion spectrum into three deontic symbols, but makes it difficult to disentangle similar opinions or preferences of (slightly) different degrees. For example, between strict vegetarians and carnivores, there are many people having different levels of preferences for meat consumption. Because these people are allowed (by their opinions) to consume both meat and vegetable[7], all of them are generally modeled in the AOI model as "omnivores" that have equal preferences for both options. Choice models have provided us with a natural solution to this issue, which is to replace the discrete deontic symbols with continuous numerical values (see Chorus et al. (2021) for an example). For example, an opinion may be described by "70% preference for meat and 30% preference for vegetable", which better suits the reality where most people are located somewhere in the opinion spectrum between the extremes. To summarize, in this avenue, I will generalize the applicability of the AOI model by endowing it with the ability to model continuous opinions.

AVENUES FOR IMPROVING THE OBFUSCATION MODEL

The obfuscation model can be regarded as an extension of the AOI model, so all the avenues for improving the AOI model are also applicable here. At the end of Chapter 4, I have explicitly mentioned the following avenues: (i) modifying model assumptions such as single-layer/ undirected networks (to multi-layer/ directed networks), positive social influence (to the mixture of positive and negative influence), and sequential updating (to parallel updating), and (ii) calibrating the model to empirical data of public opinions instead of providing illustrative examples.

AVENUES FOR IMPROVING MOBEP FRAMEWORK

At the end of Chapter 5, two avenues for future research have been proposed to improve the usability of the framework: (i) developing libraries for its components, and (ii) implementing the framework in a web-based application.

6.3.2. UTILIZING THE PROPOSED METHODS

The value of a method can be appreciated only when being used. The examples in Chapter 4 and 5 provide demonstrations of using these methods in opinion polarization studies. An intuitive and potentially fruitful starting point for such studies is to transform classic models where opinions are "what social influence influences" to models where opinions

[7] Here I have adopted a similar setting as in Section 4.4.1 of Chapter 4, where everyone only has two options: either meat or vegetable.

are private and unobservable, serving as a shortcut to constructing the initial and basic knowledge structure for this inadequately explored field. The transferring will benefit from the methods presented in this thesis, and I also expect that a cornucopia of new methods and tools will be generated during the process.

Transforming opinions from visible spins to unobservable attributes of agents is a significant step in approaching a more realistic and specific understanding of opinion polarization, but it is neither the only nor the last step as there are other features that distinguish opinions in models and in real life. For example, when modeling the dynamics of an individual's opinion, scholars usually focus on the influences from other people (i.e., social influence), but constantly ignore the effects of the individual's own attributes, such as his/her personality, experience, memory, and ideology. The obfuscation model is an ideal stepping stone towards investigating this topic as it provides an elegant way to incorporate agents' non-opinion attributes (here: agents' strategy of disclosing opinions) in opinion dynamics.

Meanwhile, as a complex social process, the dynamics of a single opinion dimension (i.e., opinion of a single topic) is hardly a closed system but usually intertwined with others. This requires us to carefully model the interplay between opinion dimensions that usually go hand-in-hand, such as abortion and gun rights (Johnson, 1997), or COVID-19 disbelief and anti-vaccine (Bok et al., 2021). Unlike models of continuous (e.g., Martins, 2008) or binary opinions (e.g., Banisch & Olbrich, 2019), the AOI model is able to model the co-evolution of relevant opinions dimensions, especially when the deontic symbols are replaced with numerical values (see Section 6.3.1). At the same time, ESBGM can provide reliable assessments of the results given that it applies to multi-dimensional data.

Furthermore, the rapid development and widespread use of information technology – including artificial intelligence, big data, and machine learning – are substantially reshaping the dynamics of opinions, which means our opinions also receive influences from external factors, such as social bots, recommendation systems, and online personalization that are driven by these new technologies. Investigating opinion dynamics in the era of information technology has obtained remarkable attention (e.g., Dandekar et al., 2013; Mäs & Bischofberger, 2015; Stella et al., 2018; Ross et al., 2019; Perra & Rocha, 2019; Keijzer & Mäs, 2021), but additional efforts are still in need to keep pace with new realities and trends. These technologies also need "behaviors" to influence opinion dynamics. For instance, social bots will mimic human behaviors like replying to messages, and recommendation systems will display personalized information to their users. Therefore, the topic falls within the scope of "opinion-behavior co-evolution", where the MOBEP framework can help. In the case of social bots, one can consider both social bots and human users as agents, and assume that social bots have fixed opinions while human users have changeable opinions that are open to social influences. The behaviors of agents are not only the messengers of opinions but also indicators of their identities (social bots or humans). The MOBEP framework can help translate these assumptions into opinion dynamics models and investigate the influence of social bots on opinion polarization.

6.3.3. DEVELOPING NEW METHODS: WITH AN EXAMPLE OF NETWORK-BASED POLARIZATION MEASUREMENT

Opinion polarization is an enormous topic with dark corners that need to be enlightened. The four methods presented in the thesis shed light on some of them, but there are more to be explored. Developing new methods based on these proposed methods is a shortcut to further explorations, as the latter have provided solid theoretical bases that promise a relatively high success rate. Different from the improved versions (See Section 6.3.1), new methods are supposed to be less directly connected with the original ones, whose core ideas, rather than their technique details, should be valued most. In the rest of this subsection, I will show one example of such new methods: a conflict-oriented network polarization measurement inspired by ESBGM.

New methods shall approach currently unanswerable questions. In the domain of polarization measurement, probably the most fundamental question is how to incorporate the "conflict nature" of polarization, which most likely originates from Esteban and Ray's claim that polarization is a "relevant correlate of potential or open social conflicts" (Esteban & Ray, 1994). Given this nature, measuring polarization becomes meaningless if no connections (and hence no conflicts) exist between people (Guerra et al., 2013). Distribution-based measurements (including all the measurements in Chapter 2) can stay away from this issue because they don't feature any network strictures (alternatively, we can say they implicitly assume that everyone is connected with everyone else), which leaves us with network-based measurements as the only option to incorporate this nature. The problem is, in network-based measurements, network edges usually don't represent connections ("I know you") but indicate endorsement ("I like you/ your opinion"), making it also incompatible with the "conflict nature".

In short, the question can be answered by a network-based polarization measurement that satisfies the following two properties:

(1) Interpersonal edges represent connections rather than endorsement; and

(2) The heterogeneity (in terms of the attribute(s) of interest) between disconnected people doesn't contribute to the polarization level.

Concerning the second property, it is natural to ask, if two previously disconnected people are now connected, will polarization go up? If we consider the connection as a channel where conflicts occur, then an additional connection between people of different groups (here: endogenously emerging groups, see Section 2.2 of Chapter 2 for details), which is an extra between-group edge in the network, shall increase polarization since it gives two enemies the chance to fight. If the new connection is between members of the same group (i.e., a new within-group edge), the story is more complex. One may argue that the new connection creates the possibility of cooperation to fight the other group, and polarization shall decrease (hereafter referred to as opinion Approach *A*). Others may argue that the new connection also offers the possibility of potential conflicts between the same-group members as long as they are not identical, which indicates that polarization should go up (hereafter referred to as Approach *B*).

ESBGM inspires us to deal with both approaches. By applying ESBG based on crisp clustering (such as K-means) to the data without considering its network structure, we

should obtain two groups of equal sizes. Different from ESBGM, now only the heterogeneity between connected pairs of data objects from the same (different) group(s) will be added to the total within- (between-) group heterogeneity. In this way, if the measurement decreases (increases) with within- (between-) group heterogeneity, additional within- (between-) group edges will always decrease (increase) polarization because they increase the within- (between-) group heterogeneity, and this measurement should then meet the requirement of Approach *A*. Meanwhile, by applying ESBG based on fuzzy clustering (see Section 6.3.1), each data point (i.e., nodes) i ($i = 1, 2, ..., N$, where N is the size of the population) should be given a clustering coefficient p_{ik} ($k = 1, 2$), which is the possibility that i is in group C_k. Given all p_{ik} for all i and k, we can generate a collection of all possible partitions of the data sets (together with their occurrence rate) that are in line with the clustering coefficients. To give the simplest possible example, if we have two nodes $i = 1, 2$, and $p_{11} = 0.5$ and $p_{21} = 1$, then in order to preserve these coefficients, we need at least two partitions with equal occurrence rate of 0.5: (i) both nodes are in C_1, and (ii) node 1 in C_2 and node i in C_1. The measurement should be the average between-group heterogeneity of all possible partitions weighted by their occurrence rate. Therefore, all edges (connections) contribute positively to polarization, but those who are considered as within-group edges in most occurrences (which means they are most likely to be classified as within-group edges in Approach *A*) will contribute less than those being considered as between-group connections in most occurrences (which means they are most likely to be classified as between-group edges in Approach *B*), and hence this measurement will be suitable for Approach *B*.

This ESBGM-inspired network-based measurement is far from completion, but it shows how the methods proposed in the thesis can inspire new methods for new research questions. After all, the development of the four proposed methods also benefits from existing ones like the K-means clustering and the voter model, and only by this type of "innovation chain" can the frontiers of the field be extended.

6.4. SOCIETAL RELEVANCE

In the introductory chapter, I have commented on the King's speech by pointing out that polarization is more than a tone of speech but a dreadful reality. Although I do not fully agree with His Majesty's opinion on polarization, I do utterly support his assertion that "[i]f every time is a time of transition, social change is a constant."[8] The year of 2021 acts as a live proof of this assertion: at the time of writing Chapter 1 (which was around mid-September), the number of daily COVID-19 cases in the Netherlands had been well maintained at a relatively low level due to the growing vaccination coverage and possibly the implementation of curfews and lockdowns. A few months later, when people were preparing for a cheering holiday season (compared to 2020), a possibly more infectious variant – Omicron – started to burst into prominence, making most optimistic predictions awfully hilarious. Opinions are even more mysterious than viruses, as there is neither a

[8] This is the official translation of the following (in Dutch): "*Als elke tijd overgangstijd is, is maatschappelijke verandering een constante.*" The full speech is available at https://www.koninklijkhuis.nl/documenten/toespraken/2021/09/21/troonrede-2021 (Dutch) and https://www.royal-house.nl/documents/speeches/2021/09/21/speech-from-the-throne-2021 (English, where the quotation in the main text is taken).

PCR test to identify one's hidden opinion, nor a dose of vaccine to slow down the spread of harmful ideas. Winston Churchill, who lost the 1945 British election as a war hero, and Hillary Clinton, who was expected to become the first female President of the United States by most polls in 2016, should agree with us with no hesitation.

So, if our societies, where opinions play a key role, simply hate pre-written scripts, why bother modeling them in the first place? It turns out that "all models (of opinion dynamics) are wrong, but some are useful"[9] in the sense that models developed in the ivory tower can actually help mitigate some problems – such as opinion polarization – that emerge during social changes, and the methods provided in this thesis could make this process easier.

In November 2021, YouTube announced that dislike counts would become invisible to viewers with the excuse that this change would "create an inclusive and respectful environment"[10] for creators. This announcement had caused widespread anger from both creators and viewers, questioning YouTube's intention and worrying about its consequence, such as whether the removal of dislike counts will alleviate or aggravate online opinion polarization. Luckily the methods of this thesis can help solve this problem as we can frame it as the opinion-behavior dynamics since most viewers' opinions are expressed by clicking the like or dislike bottom. Therefore, it will be very promising to use the AOI model, the MOBEP framework, and the polarization measurement ESBGM to study the opinion dynamics and its resulting polarization level before and after the removal of dislike counts. Furthermore, the removal is indeed an example of obfuscating behaviors, so the obfuscation model will also help capture the subtle changes in the minds and actions of both viewers and creators initiated by this controversial change. Although I don't have the space or time to include such a study in this very last chapter, the above description is already a good advertisement for using the methods of the thesis to mitigate real-life opinion polarization issues.

Over the past decades, battling opinion polarization remains a cheap talk or easy promise from politicians and business magnates who may actually benefit from opinion polarization. With research methods like those presented here, every stakeholder – from governments to ordinary citizens – can find their places in the battle as anything that is potentially relevant to opinion polarization can be formally analyzed (by scholars using these methods) and thereby practical suggestions – such as "establishing/ abolishing Internet real-name system" for policy makers, "keeping/ removing the dislike count" for social media companies, and "allowing/ rejecting cookies" for Internet users – will then be given. After all, opinions are private and unobservable, so everyone should be responsible for his/her own opinions.

6.5. REFLECTIONS

This section introduces my reflections on the research of this thesis. They are not directly related to my research goals, but may inspire and guide future studies of opinion dynamics and polarization.

[9] The quotation "all models are wrong, but some are useful" is a popular aphorism among modelers, which can be traced back to George Box, see https://www.lacan.upc.edu/admoreWeb/2018/05/all-models-are-wrong-but-some-are-useful-george-e-p-box/.

[10] See https://blog.youtube/news-and-events/update-to-youtube/ (accessed on 4 December 2021).

6.5.1. OPINION DYNAMICS VERSUS OPINION POLARIZATION

If the reader is not familiar with the topic of the thesis, he/she may be confused about two terms that often go hand-in-hand, namely "opinion dynamics" and "opinion polarization". The relation or difference between these two terms may be the first thing in the field that many people always wanted to know but were afraid to ask. This situation becomes more confusing when these two terms are used interchangeably in many papers.

If we keep a distance from the mixed usages, the difference is actually quite clear. Simply from its name, we know that "opinion dynamics" refers to the "dynamics of opinions", or "changes in opinions". Du et al. (2017) give a more specific definition that opinion dynamics is "the process in which agents form and update their opinions over time". Besides, the term "opinion dynamics" also serves as the name of a huge collection of models that describe this process ("opinion dynamics models"), as well as the name of the field where these models live ("the field of opinion dynamics").

Opinion polarization is a property of opinion distributions[11], whose meaning may differ across studies (see Chapter 2 for a comprehensive review and discussion). In other words, opinion polarization characterizes the outcome of the opinion dynamics that happened before. Traditionally, scholars focus on three types of outcomes generated by opinion dynamics, namely consensus (everyone has the same opinion), diversity (a large number of different opinions coexist), and polarization. I argue that these three outcomes can be all united under the name of polarization: consensus is the state of minimum polarization, and diversity is associated with medium levels of polarization. I can thereby draw the conclusion that every opinion dynamics model is somehow related to opinion polarization since they will end up in either consensus (low polarization), diversity (medium polarization), or (high) polarization. As a result, the boundary between opinion dynamics studies and opinion polarization studies is blurred, and this could be one of the reasons that are responsible for readers' confusion.

The above reflection provides us with new thoughts about opinion polarization research. As argued above, every outcome of an opinion dynamics model is related to opinion polarization. While a model that generates opinion polarization tells us why/how it happens, a model that generates consensus or diversity tells us how it can be avoided, which is indeed what the public and policy makers are expecting from opinion polarization studies. In this sense, we should not limit ourselves to the traditional approach that focuses on the search for models that lead to opinion polarization. Instead, equal attention should be given to other models in order to study how to promote consensus and diversity.

6.5.2. NEED FOR RIGOROUS DEFINITIONS OF COMMONLY KNOWN CONCEPTS

Human languages are vague. Expressed in the same words, what different people really mean can be very different, and this is how misunderstanding, distrust, and conflict start. As scholars are also humans, it is not surprising to find out that just like everyone else, they may use the same term while referring to different things. A more common phenomenon is that the scholarly usage of a term is different from its everyday usage, keeping relevant studies in an ivory tower isolated from the general public and policy makers.

[11] There are studies that consider opinion polarization as a property of the process of opinion dynamics (Dandekar et al., 2013), but this type of view is less popular.

This thesis has witnessed such vagueness in opinion dynamics research, which in fact motivated the whole thesis (see Chapter 1 for details). The most obvious example is about the term "opinion". Flache et al. (2017) claim that "opinion" generically refers to "what social influence influences", including concepts that are different from "opinion" such as "belief", "behavior", and "attitude". Meanwhile, in the models where opinion dynamics is coupled with behaviors (e.g., Martins, 2008), "opinion" no longer represents "behavior" and requires a less generic definition. Outside the field, the general public may feel awkward when seeing opinions being calculated and analyzed as numerical variables and observable properties of agents. It turns out that there is no consensus regarding the meaning of "opinion" whether within academia or between academia and the general public, which would lead to a series of chaotic consequences such as the hampering of knowledge accumulation and the misunderstanding between scholars and the public. Similar examples can be found in studies of opinion polarization, where "polarization" is usually considered as common knowledge that requires no explicit explanation, although people actually have very different understandings of this term (see Chapter 2).

The problem described above is caused by the lack of rigorous definitions of these concepts. Unlike natural or mathematical science, social science usually concerns concepts that are commonly known in our everyday life, such as "opinion" and "polarization". Scholars are likely to overlook the need for rigorous definitions of these concepts because they would assume that readers already know them. However, readers and even the scholars themselves may only have unclear and different understandings of these concepts. As a result, everyone interprets the concepts in their own ways, and it becomes difficult to tell if studies that carry the same name are actually talking about the same thing, and if the public has been correctly informed. To solve this problem, I would like to call for scholarly attention to the need for rigorous definitions of commonly known concepts in social science research in order to formally translate concepts in daily life into concepts in scientific contexts.

6.5.3. COMPLEXITY AND SIMPLICITY IN OPINION DYNAMICS MODELS

All models are simplifications of reality, but some models are simpler than others. In this thesis, models of different levels of complexity have been proposed. The simplest one should be the AOI model presented in Chapter 3, whose core is no more than a generalized voter model. The most complex model lies in the case study of mask wearing presented in Chapter 5, which is a combination of the conformity model (Buechel et al., 2015), rejection model (Flache & Macy, 2011), and beta-binomial model (Murphy, 2012; Khalvati et al., 2016, 2019), with modifications inspired by many other studies. As a case study that aims to illustrate how the framework – the main contribution of this chapter – works, this model has already cut off many other possibilities of modeling (e.g., modeling informational influence by a positive social influence model instead of a rejection model), which further indicates that a formal opinion dynamics model can be even more complex than the most complex one in this thesis.

It does not surprise us that opinion dynamics models have various complexity levels, but I am curious if there is any guiding philosophy about choosing the complexity level for model design. This is a non-trivial issue as everyone loves simplicity while everyone also wants a model that is close to reality. A few decades ago when scholars first attempted

to model opinion dynamics, they had few materials (e.g., modeling techniques, existing models, theories, etc.) at hand, which forced them to come up with simple and elegant models such as the voter model and DeGroot model. At our current moment when the field has been far from its infancy, staying simple becomes more difficult. On the one hand, most of the basic topics have been explored, and the remaining research questions, such as opinion polarization, are not likely to be answered by simple models. On the other hand, existing literature has warned us that certain types of model setups (e.g., modeling opinion as an observable feature of agents, not considering social network effects) are oversimplified, and more complex replacements are needed. Meanwhile, nobody wants to read a complex model whose description may occupy four or five pages: when you have (painfully) arrived at page four, you may have already forgotten what's happening on page one. In addition, unless the model is carefully built upon empirical data, an unnecessarily complex model usually means narrow applicability whose conclusions can hardly be generalized. Therefore, for those who want to study opinion dynamics in a realistic but general (i.e., not limited to particular real-life conditions such as location, topic, etc.) manner, it is crucial to find a balance between complexity and simplicity.

Let's first take a look at the simplest model in the thesis – the AOI model, whose aim is to introduce the notion of "learning opinions by observing actions" to opinion dynamics instead of replicating real-life situations. Therefore, for the AOI model, the priority is to formally and theoretically model the microscopic mechanism of how opinions can be learned by observing actions (i.e., the action-opinion inference process), while other model components that are not directly related to the notion, such as network structures and dimensionality of opinions, are given much less attention: in the original model, I simply choose the Von Neumann neighborhood as the network (which is somewhat a default choice for agent-based/ opinion dynamics models), and intentionally ignore the empirical value of modeling multi-dimensional opinions (see Chapter 2 for a discussion about uni- and multi-dimensional opinion polarization). Doubtlessly, these model components are also generally important in modeling opinion dynamics, but they are not the main characters of the AOI model, and thereby choosing anything different from the default option (which is also the simplest in most cases) will unnecessarily complicate the model. However, one of the necessary and promising directions for future studies is to test the AOI model in different social networks and opinion dimensions to see if the conclusion obtained in the default setting (Von Neumann neighborhood and uni-dimensional opinion) is still valid. To summarize, the notion or concept of interest should be first introduced in the simplest possible manner, and more complexity shall be added.

On the contrary, the most complex model in the thesis – the mask-wearing model presented in Chapter 5 – specifies each model component carefully in order to approach the real situation. For example, the small-world network used in the model resembles the limited offline interactions between humans enduring lockdowns, and the co-existence of behavior dynamics and direct opinion dynamics resembles our hybrid life where online communications on social media and offline (non-verbal) interactions in local stores are mixed. To generalize, in a model whose aim is to give an accurate and detailed description of relevant social dynamics or phenomena, all model components are of equal importance as any unrealistic or oversimplified assumption about any component will harm the reliability of the model. Therefore, I cannot assign default options to the

model components as I did in the AOI model but need to carefully specify each of them based on either theories or real situations.

To conclude, I argue that the level of complexity of an opinion dynamics model should depend on its aim. If the model is used to introduce a new notion, concept, or assumption, other model components should be kept at a minimum level of complexity if they are not directly related to the implementation of the newly introduced subject. Examples include the AOI model in Chapter 3 and the obfuscation model in Chapter 4. If the model is used to describe the reality, all model components should be carefully specified, leading to a more complex but more realistic model such as the mask-wearing model presented in Chapter 5.

6.5.4. SOME THOUGHTS ON THE FIELD OF OPINION DYNAMICS

The field of opinion dynamics[12] is unique. You can find relevant research papers from both physics journals such as Europhysics Letters and sociology journals such as the Journal of Mathematical Sociology. From the physics side, opinion dynamics offers the opportunity to apply modeling techniques that were originally designed for particles and spins to human agents, which is a challenging but exciting journey with the expectation of witnessing the hidden similarity between physics systems and human societies. From the sociology side, opinion dynamics represents a new approach to understanding social phenomena, where sociologists can control the entire system as a simple and clean representation of our society and draw conclusions by simulating or analytically solving it. With the joint contribution from physicists and sociologists, the field has endured a rapid process of development in the last decades, resulting in a large and growing number of publications.

The increase in quantity does not necessarily indicate an improvement in quality. In general, there is a tendency that the most influential papers are always these decades-old ones. There is nothing wrong with an individual model that is built upon a classic masterpiece, but for a fast-growing field, this tendency, while endorsing these classic papers, leads to the question that if the development of the field has reached a plateau.

It may be premature to make any assertion at this moment, but undoubtedly new directions are in urgent need to drive the field out of the accumulation of models (instead of contributions) to the next stage. This thesis underlies one of the new directions that take into consideration the "private and unobservable" feature of opinions by introducing behaviors to opinion dynamics. This direction dives deeper into modeling opinion-opinion interactions between agents, while other directions may be devoted to different aspects or granularity levels of opinion dynamics. For example, one ongoing trend is to study the formation of opinions within each agent by breaking opinions into smaller components such as arguments (Mäs & Flache, 2013; Feliciani et al., 2021) and attitude elements (van der Maas et al., 2020). Furthermore, by integrating opinion formation with argumentation frameworks, we will be able to model deliberations inside an agent and between agents (Taillandier et al., 2019), implying that this approach is potentially useful in explaining opinion polarization in a more realistic and detailed manner (see Proietti (2017) for a minimalistic example).

[12] To avoid unnecessary complexity, here, "the field of opinion dynamics" also includes the field of opinion polarization. See Section 6.5.1 for a discussion.

Apart from these particular directions, the future of this interdisciplinary field is most likely to be driven by interdisciplinary forces. One of the driving forces comes from the techniques that originated from physics and mathematics, such as the hierarchical Ising model (van der Maas et al., 2020), random field Ising model (Tiwari et al., 2021), matrix-weighted Laplacian dynamics (Ahn et al., 2020), and sheaf model (Hansen & Ghrist, 2021). The adoption of these new techniques allows us to investigate a wide range of new topics, expanding the territory of opinion dynamics studies. The second driving force comes from empirical data. A common criticism of the field lies in the lack of empirical support. As a response, a growing number of opinion studies use empirical data, whether from surveys, field studies, or lab experiments, to calibrate and validate their models, or use opinion dynamics models (with simulations) to test, explain, and generalize the empirically observed findings (see Kozitsin (2021), and Keijzer & Mäs (2021) for some very recent examples). In the integration between opinion dynamics research, physics techniques, and empirical studies lies the next frontier of computational sociology, where the polarization of the "private and unobservable opinions" is likely to be better understood.

REFERENCES

[1] Aggarwal, C. C., & Reddy, C. K. (2014). Data clustering: Algorithms and applications. CRC Press.

[2] Ahn, H. S., Van Tran, Q., Trinh, M. H., Ye, M., Liu, J., & Moore, K. L. (2020). Opinion dynamics with cross-coupling topics: Modeling and analysis. IEEE Transactions on Computational Social Systems, 7(3), 632-647.

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

[4] Bishop, C.M (2006).Pattern recognition and machine learning. Springer.

[5] Bok, S., Martin, D. E., & Lee, M. (2021). Validation of the COVID-19 Disbelief Scale: Conditional indirect effects of religiosity and COVID-19 fear on intent to vaccinate. Acta Psychologica, 219, 103382.

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

[7] Chawla, S., & Gionis, A. (2013). k-means–: A unified approach to clustering and outlier detection. In Proceedings of the 2013 SIAM International Conference on Data Mining (pp. 189-197). Society for Industrial and Applied Mathematics.

[8] Chorus, C., van Cranenburgh, S., Daniel, A. M., Sandorf, E. D., Sobhani, A., & Szép, T. (2021). Obfuscation maximization-based decision-making: Theory, methodology and first empirical evidence. Mathematical Social Sciences, 109, 28-44.

[9] 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).

[10] 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.

[11] Dong, Y., Zhan, M., Kou, G., Ding, Z., & Liang, H. (2018). A survey on the fusion process in opinion dynamics. Information Fusion, 43, 57-65.

[12] Du, E., Cai, X., Sun, Z., & Minsker, B. (2017). Exploring the role of social media and individual behaviors in flood evacuation processes: An agent-based modeling approach. Water Resources Research, 53(11), 9164-9180.

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

[14] Feliciani, T., Flache, A., & Mäs, M. (2021). Persuasion without polarization? Modelling persuasive argument communication in teams with strong faultlines. Computational and Mathematical Organization Theory, 27(1), 61-92.

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

[16] 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).

[17] Guerra, P. C., Meira Jr, W., Cardie, C., & Kleinberg, R. (2013). A measure of polarization on social media networks based on community boundaries. In Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (pp. 215 - 224).

[18] Guha, S., Rastogi, R., & Shim, K. (1998). CURE: An efficient clustering algorithm for large databases. In Proceedings of ACM SIGMOD International Conference on Management of Data (pp. 73 – 84).

[19] HajKacem, M.A.B., N'Cir, CE.B., & Essoussi, N. (2019). Overview of scalable partitional methods for big data clustering. In O. Nasraoui & CE. Ben N'Cir (Eds.), Clustering methods for big data analytics. Springer.

[20] Hansen, J., & Ghrist, R. (2021). Opinion dynamics on discourse sheaves. SIAM Journal on Applied Mathematics, 81(5), 2033-2060.

[21] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651-666.

[22] Johnson, N. J. (1997). Principles and Passions: The Intersection of Abortion and Gun Rights. Rutgers Law Review, 50, 97-197.

[23] Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7), 881-892.

[24] Kaufman, L., & Rousseeuw, P. J. (1990). Finding groups in data: An introduction to cluster analysis. John Wiley & Sons.

[25] Keijzer, M. A., & Mäs, M. (2021). The strength of weak bots. Online Social Networks and Media, 21, 100106.

[26] 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).

[27] 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).

[28] Kozitsin, I. V. (2021). Opinion dynamics of online social network users: A micro-level analysis. The Journal of Mathematical Sociology. https://doi.org/10.1080/0022250X.2021.1956917

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

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

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

[32] Masad, D., & Kazil, J. (2015). MESA: An agent-based modeling framework. In Proceedings of the 14th Python in Science Conference (pp. 51-58).

[33] Mobilia, M. (2003). Does a single zealot affect an infinite group of voters?. Physical Review Letters, 91(2), 028701.

[34] Müller, A.C., & Guido, S. (2006). Introduction to machine learning with Python: A guide for data scientists. O'Reilly Media.

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

[36] Olukanmi, P. O., & Twala, B. (2017). K-means-sharp: Modified centroid update for outlier-robust k-means clustering. In 2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech) (pp. 14-19).

[37] Pelleg, D., & Moore, A. (1999). Accelerating exact k-means algorithms with geometric reasoning. In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 277-281).

[38] Perra, N., & Rocha, L. E. C. (2019). Modelling opinion dynamics in the age of algorithmic personalisation. Scientific Report, 9(1), 7261.

[39] Proietti, C. (2017). The Dynamics of group polarization. In A. Baltag, J. Seligman & T. Yamada (Eds.), Logic, rationality, and interaction (LORI 2017 Proceedings) (pp. 195-208), Springer.

[40] 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.

[41] Stella, M., Ferrara, E., & De Domenico, M. (2018). Bots increase exposure to negative and inflammatory content in online social systems. Proceedings of the National Academy of Sciences, 115(49), 12435-12440.

[42] Stoffel K., & Belkoniene A. (1999). Parallel k/h-Means clustering for large data sets. In Proceedings of Euro-Par'99 Parallel Processing (pp. 1451-1454). Springer.

[43] Taillandier, P., Salliou, N., & Thomopoulos, R. (2019). Coupling agent-based models and argumentation framework to simulate opinion dynamics: Application to vegetarian diet diffusion. Social Simulation Conference 2019. Mainz, Germany. https://doi.org/10.3929/ethz-b-000383850

[44] Tiwari, M., Yang, X., & Sen, S. (2021). Modeling the nonlinear effects of opinion kinematics in elections: A simple Ising model with random field based study. Physica A: Statistical Mechanics and its Applications, 582, 126287.

[45] Tsang, A., & Larson, K. (2014). Opinion dynamics of skeptical agents. In Proceedings of the 2014 International Conference on Autonomous Agents and Multiagent Systems (pp. 277-284).

[46] van der Maas, H. L. J., Dalege, J., & Waldorp, L. (2020). The polarization within and across individuals: The hierarchical Ising opinion model. Journal of Complex Networks, 8(2), cnaa010.

[47] Verma, G., Swami, A., & Chan, K. (2014). The impact of competing zealots on opinion dynamics. Physica A: Statistical Mechanics and its Applications, 395, 310-331.

[48] Waagen, A., Verma, G., Chan, K., Swami, A., & D'Souza, R. (2015). Effect of zealotry in high-dimensional opinion dynamics models. Physical Review E, 91(2), 022811.

[49] Xu, R., & Wunsch, D. C. (2009). Clustering. Wiley. https://doi.org/10.1002/9780470382776

[50] Yildiz, E., Ozdaglar, A., Acemoglu, D., Saberi, A., & Scaglione, A. (2013). Binary opinion dynamics with stubborn agents. ACM Transactions on Economics and Computation (TEAC), 1(4), 1-30.

[51] 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

SUMMARY

BACKGROUND AND RESEARCH GOALS

Polarized opinions are everywhere. From opposite attitudes towards Hawaiian pizza to the partisan divide in the United States, we have experienced enough opinion polarization in recent years. Sadly, it is usually a sign of follow-up criticism when people start to talk about "opinion polarization". The term, which should neutrally describe a widespread social phenomenon, has been proven to be associated with different dismaying outcomes, ranging from hostility to civil wars. Given its harmful consequence, few would doubt the urgent need for a solution to this long-lasting issue, and such a solution requires a deep understanding of opinion polarization in real-life situations.

The urgent need has motivated remarkable research efforts in the past few decades. Especially in the domain of computational sociology, a considerable amount of opinion dynamics models have been proposed to explain opinion polarization from microscopic mechanisms that govern interactions between individuals. A common feature of these models, which probably results from their roots in statistic physics, is that opinions are observable and can be directly affected by other opinions just like a "spin" in the famous Ising model. However, opinions in real life are of fundamental difference from "spin" in the sense that it is by nature private and unobservable, whose expression, transmission, and inference largely depend on observable behaviors: even if people are allowed to verbally exchange opinions, how these opinions are translated into words and how these words are inferred by both parties still play a critical role in the dynamics of opinions. Thereby, we could put forward a thesis (which we did, literally) that there is a fundamental discrepancy between opinion polarization in the literature and opinion polarization in real-life situations that would deteriorate our trust in these models, let alone the solutions generated accordingly.

The discrepancy naturally leads us to wonder, "*how can the polarization of private and unobservable opinions be explained?*" Indeed, this is an open question with various potential answers. Therefore, the primary research goal of the thesis is not to provide an ultimate explanation but "*to develop methods that could systematically support the exploration of mechanisms that can explain the polarization of private and unobservable opinions*". Specifically, the primary goal can be divided into four sub-goals:

- Sub-goal 1: To develop a formal and broadly applicable polarization measurement that is coherent with the notion of group.
- Sub-goal 2: To develop a modeling method that can incorporate the "private and unobservable" feature of opinions in opinion polarization mechanisms.

I am indebted to Tom van den Berg for helping me translate this Summary into Dutch.

- Sub-goal 3: To develop a modeling method to study the effect of the deviation from the "default strategy", such as obfuscation, on opinion polarization.

- Sub-goal 4: To develop a unifying framework of the co-evolution of opinion and behavior to organize existing efforts and facilitate future works of opinion polarization mechanisms.

CONTENT OF THE THESIS

The thesis centers on the four sub-goals. From Chapter 2 to 5, each chapter contributes to one of the sub-goals, and their contents are summarized as follows.

Chapter 2 presents a group-based polarization measurement, called ESBGM (**E**qual **S**ize **B**inary **G**rouping **M**easurement), that uniquely reflects the widely accepted idea that groups, instead of individuals, are the crucial actors in conceptualizing polarization. The chapter first unveils one of the major obstacles to measuring polarization, which is the discrepancy between how we understand and measure polarization: it is an article of faith among most scholars that groups shall be kept center stage in the conceptualization of polarization, but when it comes to measuring it, few know what exactly groups are. The measurement tackles this problem by employing a grouping method called ESBG (**E**qual **S**ize **B**inary **G**rouping) that divides the data set of interest into two groups of equal sizes according to data similarities. ESBG is then justified as a proper grouping method that can overcome certain theoretical and practical nuisances discovered in other grouping methods, including discontinuity issues and contradiction of axioms. Inspired by clustering techniques, the chapter implements ESBG in a similar fashion as K-means and applies it to a two-dimensional synthetic data set. At the end of the chapter, ESBGM is compared with bimodality (e.g., Ashman's D and bimodal separation) and bipolarization (e.g., Wolfson's index) measurements. The "squeezing-and-moving" framework is then developed to better explain the relation between ESBGM and bipolarity measurements. For opinion polarization studies, ESBGM functions as a reliable (in the sense that the notion of groups is well incorporated in this measurement) method to quantify results and generate conclusions.

Chapter 3 introduces an agent-based modeling method/ model, called AOI (**A**ction-**O**pinion **I**nference) model, for modeling the co-evolution of discrete opinions and discrete behaviors, with a particular focus on the multiplicity of opinion-behavior relations. The chapter starts with the notion of "*learning opinions by observing actions*", in contrast with the widely used notion of "*learning opinions by observing opinions*". The former, which is the cornerstone of the AOI model, means that people learn opinions of others by observing (and then interpreting) their behaviors, while the latter, implicitly held in most existing opinion dynamics models, assumes that opinions are directly observable. The AOI model is built upon the former notion, which is believed to be closer to reality than the latter. In the model, one's behavior (called "actions" in the chapter) is governed by her opinion, and one updates her opinion according to the behaviors she observed in her neighborhood. The behavior-opinion relation, described by deontic logic, determines how one's opinion governs her behavior, and the inference process, based on Bayesian learning, determines how the observed opinions affect one's opinion. The model is applied to different situations, leading to the conclusion that the dynamics of opinions

is largely determined by the action-opinion relations. The mathematical derivation of a particular simulation result is provided to help better understand the process. In addition, the notion of "*learning opinions by observing actions*" is compared with similar concepts such as information cascade, persuasion model, and pluralistic ignorance. The model itself is compared with the CODA model, the (constrained) voter model, and the language competition model. From the perspective of opinion polarization studies, the AOI model establishes a novel modeling method to cope with the "private and unobservable" feature of opinions, serving as one of the pioneering models of opinion-behavior co-evolution.

Chapter 4 is devoted to an agent-based modeling method/ model for exploring the effect of obfuscation – a representative of behavioral choice strategies of intelligent agents – on opinion polarization. The chapter begins with the observation that most opinion dynamics models equate hiding opinions with keeping silent, although in many cases, people use a more complex and intelligent strategy to hide their opinions without deception. The strategy, called *obfuscation*, helps people hide opinions by choosing the behavior that minimizes the disclosed information about the underlying opinion. On the basis of the AOI model, the formal opinion dynamics model of obfuscation is proposed to explore the effect of obfuscation on opinion dynamics (including opinion polarization). Two examples are given to illustrate how the model can be used to simulate obfuscation in opinion dynamics. The first example describes the simplest case where vegetarians and omnivores are asked to choose between beef and salad. Simulations imply that the popularity of exclusive opinions (opinions that only allows one behavior, which is "vegetarianism" in this case) is positively related to the percentage of obfuscators in the population. The second example replicates the story of "*Emperor's New Clothes*", and in this particular example, obfuscation is found to facilitate the spread of misinformation: the Emperor is dressed. These conclusions are valid only in the particular settings of the examples, and the generic conclusion is that the effect of obfuscation relies on the relations between opinions and behaviors. In the field of opinion dynamics/ polarization, the "default strategy" for agents is to express their opinions honestly and correctly. The model pushes forward the frontiers of opinion dynamics modeling by focusing on a typical "deviate strategy" (i.e., obfuscation) that plays a potentially crucial part in explaining opinion polarization.

Chapter 5 develops a unifying framework, called MOBEP (**M**odeling **O**pinion-**B**ehavior Co-evolution for **E**xplaining Opinion **P**olarization), of modeling opinion-behavior co-evolution for explaining opinion polarization. The chapter first acknowledges modeling opinion-behavior co-evolution as a novel approach to explaining opinion, considering the inseparability between opinions and behaviors. A considerable number of models that are concerned with opinion-behavior co-evolution are categorized and reviewed, and a central notion that "*behaviors serve as the messenger of opinions*" is derived. Based on the reviews, the MOBEP framework is developed to organize the uncoordinated works accumulated till now and facilitate opinion polarization studies in the future. The framework contains five key components (i.e., opinion-driven behavior change, normative influence, behavior-opinion inference, informational influence, and direct opinion influence) and two implementation components (i.e., schedule and polarization measurement), among which the component "behavior-opinion inference" is one of the highlights that distinguish the framework from its competitors. The framework is tested by a selection of

opinion-behavior co-evolution models: these models are decomposed according to the framework, and it is clear that the framework is able to accommodate them. As a demonstration of how the framework can facilitate future studies, a case study of mask wearing during the COVID-19 pandemic is built from scratch. For opinion polarization studies, MOBEP is not only a framework but also a call for attention to this new and promising approach, given that insights from existing models can be easily accumulated and future studies can be conducted in a more systematic way with the help of the framework.

CONCLUSIONS

In a sentence, the thesis makes methodological contributions to the field of opinion polarization by developing four research methods to systematically support the exploration of opinion polarization mechanisms under the realistic assumption that opinions are private and unobservable, requiring corresponding observable behaviors as their messengers to influence each other. Each of the methods can be used independently, but they are internally connected. The latter three methods (all except ESBGM) all center on the notion that "*behaviors are the messenger of opinions*" – a refined version of AOI's central notion of "*learning opinions by observing actions*". The first cornerstone is laid by the AOI model (Chapter 2), which provides both theoretical and technical basis for modeling opinion-behavior co-evolution under this notion. The boundaries of the AOI model are extended by the obfuscation model (Chapter 4) by diving into a particular "deviate strategy" of obfuscation. Instead of focusing on specific modeling tasks, the MOBEP framework (Chapter 5) takes a broader view and aims to help accumulate existing insights (including those derived from the other three methods) and promote future research. Finally, ESBGM (Chapter 2) provides a reliable way to measure the results generated by the other three methods, connecting them to the primary research goal of explaining opinion polarization. The four methods are hence integrated into one toolkit for almost all aspects of opinion polarization studies, with the ability to deepen our understanding of the polarization of "private and unobservable opinions", and help depolarize the real world that is now being torn apart.

SAMENVATTING

ACHTERGROND EN ONDERZOEKSDOELEN

Gepolariseerde meningen zijn overal. Of je nu denkt aan tegenovergestelde meningen over de pizza Hawaii of aan de politieke verdeeldheid in de Verenigde staten, in de afgelopen jaren waren gepolariseerde meningen nooit ver weg. Helaas, is het vaak een teken van vervolgkritiek wanneer men spreekt over "opiniepolarisatie". Onderzoek laat zien dat de term, die een wijdverbreid sociaal fenomeen neutraal probeert te beschrijven, is geassocieerd met verschillende schadelijke fenomenen, variërend van vijandigheid tot aan burgeroorlogen. Gegeven de schadelijke gevolg, twijfelen maar weinig mensen aan de noodzaak om tot een oplossing te komen voor dit langdurige probleem. Hiervoor is een diepgravend begrip nodig van opiniepolarisatie in werkelijke omstandigheden.

Deze dringende behoefte heeft de afgelopen decennia tot veel onderzoeksinspanningen geleid. Vooral binnen het domein van de computationele sociologie is er een groot aantal opiniedynamiekmodellen ontwikkeld om opiniepolarisatie te verklaren vanuit de microscopische mechanismen die de interacties tussen personen bepalen. Een gedeeld kenmerk van deze modellen, die waarschijnlijk voortvloeit uit hun wortels in de statistische fysica, is dat meningen als waarneembaar worden voorgesteld en direct kunnen worden beïnvloed door andere meningen, vergelijkbaar met een "spin" in het beroemde Ising-model. Echter, meningen in het echte leven verschillen fundamenteel van een dergelijke "spin" omdat zij van nature privé en niet waarneembaar zijn. De uitdrukking, overdracht en afleiding van werkelijke meningen hangt grotendeels af van waarneembaar gedrag: zelfs in het geval dat mensen verbaal meningen uitwisselen, dan nog speelt de manier waarop meningen worden vertaald in woorden en hoe deze woorden vervolgens door beide partijen worden afgeleid een cruciale rol in de dynamiek van meningen. Op basis hiervan kunnen we de these formuleren (en dat hebben we ook letterlijk gedaan) dat er een fundamenteel verschil zit tussen de opiniepolarisatie zoals die binnen de literatuur wordt beschreven en werkelijke gevallen van opiniepolarisatie. Dit tast de betrouwbaarheid van de beschreven modellen aan, alsook die van de hierop gebaseerde oplossingen.

Deze discrepantie brengt ons tot de vraag: "*hoe kan de polarisatie van private en niet-waarneembare meningen worden verklaard?*" Dit is een open vraag waarop verschillende antwoorden mogelijk zijn. Het primaire onderzoeksdoel van dit proefschrift is daarom niet om een ultieme verklaring te geven, maar om "*methoden te ontwikkelen om op een systematische manier het onderzoek naar mechanismen te ondersteunen die de polarisatie van private en onwaarneembare meningen verklaren".* Concreet kan dit primaire doel worden onderverdeeld in vier subdoelen:

- Subdoel 1: Het ontwikkelen van een formele en breed toepasbare polarisatiemeting die de notie van "groep" kan incorporeren.

- Subdoel 2: Het ontwikkelen van een modelleringsmethode die de "private en niet-waarneembare" eigenschap van meningen in opiniepolarisatiemechanismen kan opnemen.
- Subdoel 3: Het ontwikkelen van een modelleringsmethode om het effect te bestuderen van het afwijken van de "standaardstrategie" van opiniepolarisatie, zoals verhullen (obfuscation).
- Subdoel 4: Het ontwikkelen van een verenigend raamwerk van de co-evolutie van opinie en gedrag om bestaande resultaten te organiseren en toekomstige studies van opiniepolarisatiemechanismen te vergemakkelijken.

INHOUD VAN HET PROEFSCHRIFT

Het proefschrift richt zich op de genoemde vier subdoelen. Van hoofdstuk 2 tot en met 5, behandelt ieder hoofdstuk er één. Hieronder volgt hiervan de samenvatting. *Hoofdstuk 2* presenteert een groepsgebaseerde polarisatiemeting, genaamd ESBGM (**E**qual **S**ize **B**inary **G**rouping **M**easurement), die op unieke wijze het algemeen aanvaarde idee weerspiegelt dat groepen, in plaats van individuen, de cruciale actoren zijn bij het conceptualiseren van polarisatie. Het hoofdstuk zet eerst uiteen dat één van de belangrijkste obstakels voor het meten van polarisatie de discrepantie is tussen hoe we polarisatie begrijpen en meten: het is een speerpunt van de meeste onderzoekers dat groepen centraal moeten worden gesteld in de conceptualisering van polarisatie, maar als het vervolgens wordt gemeten, weten maar weinigen wat groepen precies zijn. De meting lost dit probleem op door gebruik te maken van een groeperingsmethode genaamd ESBG (**E**qual **S**ize **B**inary **G**rouping), die de gebruikte dataset verdeelt in twee groepen van gelijke grote, aan de hand van dataovereenkomsten. Op deze manier is ESBG een goede groeperingsmethode die bepaalde theoretische en praktische problemen kan overwinnen, die in andere groeperingsmethoden zijn ontdekt, waaronder problemen met discontinuïteit en tegenstrijdigheid van axioma's. Geïnspireerd door clustertechnieken, implementeert dit hoofdstuk ESBG op een vergelijkbare manier als K-means en past het toe op een tweedimensionale synthetische dataset. Aan het einde van het hoofdstuk wordt ESBGM vergeleken met bimodaliteit (bijv. Ashman's D en bimodale scheiding) en bipolarisatie (bijv. Wolfson's index) metingen. Vervolgens wordt het "squeezing-and-moving" raamwerk ontwikkeld om de relatie tussen ESBGM en bipolariteitsmetingen beter te verklaren. Voor opiniepolarisatieonderzoeken fungeert ESBGM als een betrouwbare methode (in de zin dat het begrip "groepen" goed is opgenomen in deze meting) om resultaten te kwantificeren en conclusies te trekken.

Hoofdstuk 3 introduceert een agent-gebaseerd modelleringsmethode/model, genaamd het AOI (**A**ction-**O**pinion **I**nference) model, voor het modelleren van de co-evolutie van discrete meningen en discrete gedragingen, met een bijzondere focus op de veelheid van opinie-gedragsrelaties. Het hoofdstuk begint met het idee van "*meningen leren door het observeren van acties*", in tegenstelling tot het veelgebruikte "*meningen leren door meningen te observeren*". Het eerste, de hoeksteen van het AOI-model, houdt in dat mensen meningen van anderen leren door hun gedrag te observeren (en vervolgens te interpreteren), terwijl het laatste, impliciet aangenomen door de meeste bestaande modellen voor opiniedynamiek, ervan uitgaat dat meningen direct waarneembaar zijn. Het AOI-model

is gebaseerd op het eerste idee, waarvan we denken dat het dichter bij de realiteit staat dan het laatste. In het model wordt iemands gedrag (in het hoofdstuk "actions" genoemd) bepaald door haar mening, en men werkt haar mening bij op basis van het gedrag dat ze in haar buurt heeft waargenomen. De gedrags-opinie relatie, die wordt beschreven door deontische logica, bepaalt hoe iemands mening haar gedrag bepaalt. Het gevolgtrekkingsproces, gebaseerd op Bayesiaans leren, bepaalt hoe de waargenomen meningen iemands mening beïnvloeden. Het model wordt toegepast op verschillende situaties, wat leidt tot de conclusie dat de dynamiek van meningen grotendeels bepaald wordt door de gedrags-opinie relaties. Om het proces beter te begrijpen wordt de wiskundige afleiding van elk simulatieresultaat gegeven. Daarnaast wordt het idee van "*het leren van meningen door het observeren van acties*" vergeleken met gelijkaardige concepten zoals informatiecascade, het overtuigingsmodel en pluralistische onwetendheid. Het model zelf wordt vergeleken met het CODA-model, het (beperkte) kiezersmodel en het taalcompetitiemodel. Vanuit het perspectief van opiniepolarisatiestudies, stelt het AOI-model een nieuwe modelleringsmethode voor om met het "private en niet-waarneembare" kenmerk van meningen om te gaan, en kan het dienen als een van de baanbrekende modellen van co-evolutie van opinies en gedrag.

Hoofdstuk 4 is gewijd aan een op agenten gebaseerde modelleringsmethode/-model om het effect te onderzoeken van verhulling (obfuscation) – een voorbeeld van een gedragskeuzestrategieën van intelligente agenten - op opiniepolarisatie. Het hoofdstuk begint met de observatie dat de meeste modellen voor opiniedynamiek het verbergen van meningen gelijkstellen aan zwijgen, terwijl mensen in veel gevallen een complexere en intelligentere strategie gebruiken om hun mening te verbergen zonder te misleiden. De strategie, *obfuscation* genaamd, helpt mensen om hun mening te verbergen door gedrag te kiezen waarbij het onthullen van informatie over de onderliggende mening wordt geminimaliseerd. Op basis van het AOI-model, stellen we het formele opiniedynamiekmodel van obfuscation voor om het effect van obfuscation op opiniedynamiek (inclusief opiniepolarisatie) te onderzoeken. Er worden twee voorbeelden gegeven om te illustreren hoe het model kan worden gebruikt om verhulling in opiniedynamiek te simuleren. Het eerste voorbeeld beschrijft het meest eenvoudige geval waarin vegetariërs en alleseters moeten kiezen tussen rundvlees en salade. Simulaties impliceren dat de populariteit van exclusieve meningen (meningen die slechts één gedrag toelaten, in dit geval "vegetarisme") positief gerelateerd is aan het percentage obfuscators in de populatie. Het tweede voorbeeld repliceert het verhaal van "*De nieuwe kleren van de keizer*". In dit specifieke voorbeeld blijkt verhulling de verspreiding van verkeerde informatie te vergemakkelijken: de keizer is gekleed. Deze conclusies gelden alleen in de specifieke context van deze voorbeelden. De algemene conclusie is dat het effect van verhulling afhankelijk is van de relaties tussen meningen en gedrag. Binnen het gebied van opiniedynamiek/polarisatie is het de "standaardstrategie" voor agenten om hun mening eerlijk en correct te uiten. Het model verlegt de grenzen van de modellering van opiniedynamiek door zich te richten op een typische "afwijkende strategie" (d.w.z. verhulling) die een potentieel cruciale rol speelt bij het verklaren van opiniepolarisatie.

Hoofdstuk 5 ontwikkelt een verenigend raamwerk, genaamd MOBEP (**M**odeling **O**pinion-**B**ehavior Co-evolution for **E**xplaining Opinion **P**olarization), van het modelleren van de co-evolutie van opinie en gedrag voor het verklaren van opiniepolarisatie. Het hoofdstuk

erkent eerst het modelleren van co-evolutie van opinies en gedrag als een nieuwe benadering voor het verklaren van meningen, gezien de onscheidbaarheid van meningen en gedragingen. Een aanzienlijk aantal modellen die zich bezighouden met co-evolutie van opinies en gedrag worden behandeld en gecategoriseerd, en hieruit wordt de centrale notie afgeleid dat "*gedragingen dienen als de boodschapper van meningen*". Op basis van de review is het MOBEP-raamwerk ontwikkeld om de studies die tot nu toe zijn gedaan te organiseren en opiniepolarisatiestudies in de toekomst te vergemakkelijken.

Het raamwerk bevat vijf hoofdcomponenten (op opinie gebaseerde gedragsverandering, normatieve invloed, gedragsopinie-inferentie, informatie-invloed en directe opinie-invloed) en twee implementatiecomponenten (planning en polarisatiemeting). Het "gedrags-opinie-inferentie"-component maakt dit raamwerk onderscheidend ten op zicht van andere kaders. Het raamwerk is getest door middel van een selectie van opinie-gedrag co-evolutiemodellen: deze modellen worden volgens het raamwerk ontbonden, en het is duidelijk dat het raamwerk ze kan accommoderen. Ter demonstratie van hoe het raamwerk toekomstige studies kan vergemakkelijken, is een casestudy van het dragen van maskers tijdens de COVID-19-pandemie vanaf het begin opnieuw opgebouwd. Voor opiniepolarisatiestudies is MOBEP niet alleen een raamwerk, maar ook een oproep om meer aandacht te besteden aan deze nieuwe en veelbelovende aanpak, aangezien met behulp van dit kader inzichten uit bestaande modellen eenvoudig kunnen worden verzameld en toekomstige studies op een systematischere manier kunnen worden uitgevoerd.

CONCLUSIES

In één zin levert het proefschrift methodologische bijdragen op het gebied van opiniepolarisatie, door vier onderzoeksmethoden te ontwikkelen, die op een systematische manier het onderzoek naar opiniepolarisatiemechanismen ondersteunen, onder de realistische veronderstelling dat meningen privé en niet-waarneembaar zijn, waarbij wederzijds waarneembaar gedrag vereist is als hun boodschappers om elkaar te beïnvloeden. Elk van de methoden kan onafhankelijk worden gebruikt, maar ze zijn onderling met elkaar verbonden. De laatste drie methoden (allemaal behalve ESBGM) zijn allen gebaseerd op het idee dat "*gedrag de boodschapper van meningen*" is – een verfijnde versie van AOI's centrale notie van "*meningen leren door acties te observeren*". De eerste hoeksteen wordt gelegd door het AOI-model (Hoofdstuk 2), dat zowel de theoretische als technische basis biedt voor het modelleren van de co-evolutie van meningen en gedrag. De grenzen van het AOI-model worden verlegd door het obfuscation-model (Hoofdstuk 4) door ons te verdiepen in een bepaalde "afwijkende strategie" van verhulling. In plaats van zich te concentreren op specifieke modelleringstaken, neemt het MOBEP-raamwerk (Hoofdstuk 5) een bredere perspectief en heeft het tot doel bestaande inzichten (inclusief de inzichten die zijn afgeleid van de andere drie methoden) bij elkaar te brengen en toekomstig onderzoek te bevorderen. Ten slotte, biedt ESBGM (Hoofdstuk 2) een betrouwbare manier om de resultaten van de andere drie methoden te meten, en deze te verbinden met het primaire onderzoeksdoel, namelijk het verklaren van opiniepolarisatie. De vier methoden zijn daarom geïntegreerd binnen één toolkit, voor bijna ieder aspect van opiniepolarisatieonderzoek, met het vermogen om ons begrip van de polarisatie van "privé en niet-waarneembare meningen" te verdiepen en de echte wereld, die nu uit elkaar wordt getrokken, te helpen depolariseren.

CURRICULUM VITÆ

Tanzhe TANG

RESEARCH

2021–present	Postdoctoral researcher Faculty of Behavioral & Social Sciences University of Groningen, The Netherlands.
2018–2022	Ph.D. researcher Faculty of Technology, Policy, and Management Delft University of Technology, The Netherlands.

EDUCATION

2016–2017	MSc., Non-Equilibrium Systems: Theoretical Modelling, Simulation and Data-Driven Analysis King's College London, United Kingdom.
2011–2015	Bachelor in Economics Zhejiang University, China.

ACTIVITIES

2020–2021	Supervision of master students: Stella Mulia (TU Delft).
2018–2019	Editor Assistant: European Journal of Transport and Infrastructure Research.

LIST OF PUBLICATIONS

Only publications that are related to the Ph.D. research will be mentioned here.

JOURNAL PUBLICATIONS

1. Tang, T., Ghorbani, A., Squazzoni, F., & Chorus, C. G. (2021). Together alone: A group-based polarization measurement. Quality & Quantity.
   DOI:10.1007/s11135-021-01271-y

2. Tang, T., Ghorbani, A., & Chorus, C. G. (2021). Hiding opinions by minimizing disclosed information: An obfuscation-based opinion dynamics model. Journal of Mathematical Sociology.
   DOI:10.1080/0022250X.2021.1929968

3. 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).
   DOI:10.18564/jasss.4020

INTERNATIONAL CONFERENCE PRESENTATIONS

1. Tang, T., Ghorbani, A., & Chorus, C. G. (2021). Explaining polarization by coupling dynamics of opinions and behaviors under conformity: A case study of face mask wearing during the COVID-19 pandemic. Social Simulation Conference (virtual and Cracow, Poland), September 2021.

2. Tang, T., Ghorbani, A., & Chorus, C. G. (2021). Conflict-oriented polarization measurement based on network structures. Networks 2021: A Joint Sunbelt and NetSci Conference (virtual), July 2021.

3. Tang, T., & Chorus, C. G. (2019). Understanding the learning mechanism underlying social influence-based discrete choice: An Action-Opinion Inference model. International Choices Modeling Conference (Kobe, Japan), August 2019.

4. Tang, T., Ghorbani, A., & Chorus, C. G. (2019). AOIO: Action-Opinion Inference Model with Obfuscation. The 12th Annual INAS (International Network of Analytical Sociologists) Conference (St. Petersburg, Russia), August 2019.

5. Tang, T., & Chorus, C. G. (2018). Learning opinions by observing actions: Simulation of opinion dynamics using an action-opinion inference model. The 21st International Conference on Principles and Practice of Multi-Agent Systems (Tokyo, Japan), November 2018.