THE UNIVERSITY OF CHICAGO

Self-fulfilling Stigmatization: An Evolutionary Game of Online Opinion Dynamics in Social Movements

By

Hongding Zhu

July 2022

A paper submitted in partial fulfillment of the requirements for the

Master of Arts degree in the

Master of Arts Program in the Social Sciences

Faculty Advisor: Zhaotian Luo

Preceptor: Yan Xu

Self-fulfilling Stigmatization: An Evolutionary Game of Online Opinion Dynamics in Social Movements

Hongding Zhu, University of Chicago

July 29, 2022

Abstract

To illustrate how the stigmatization of social movements arises from its microfoundations—emotional confrontation, I model the process of such stigmatization on social media using an evolutionary game with two sub-populations—the participants of a social movement and outsiders in the general public. Based on the critical assumption of "hostility as retaliation", the members of the movement, who are hostile toward the general public, and haters in the general public reproduce each other. Suppose there are enough such hostile members so that it exceeds the stigmatization threshold. Then, the social movement is stigmatized and locked in conflicts with haters outside the movement. The stigmatization of a social movement can be a self-fulfilling prophecy. When there are initially enough haters, even if a social movement is peaceful at the beginning, the speed at which the movement becomes hostile under outgroup pressure will outpace the rate at which it can win public support, and finally induces the stigmatization equilibrium.

Keywords: Social movement; stigmatization; evolutionary game; Echo Chamber; censorship.

1 Introduction

More and more social movements—from Black Lives Matter to MeToo, from the climate movement to Occupy Wall Street—are being initiated and are happening online (Hara and Huang, 2011). The ease of setting up groups, the global reach of networked communities, and the convenient facilitation of organizational activities all contribute to the widespread use of social media sites for social movements (Wall, 2007). Moreover, under the tremendous political risk of offline activism, social movements in autocracy rely on online discourse even more than their counterparts in democracy (Feng, 2017; Tan, 2017). In cyberspace, people can quickly obtain information about a social movement and interact with its participants, forming their opinions about the movement accordingly.

Stigmatization refers to that social movements are misunderstood by society as the advocated value are distorted, and the participants are tagged with negative characteristics. We see significant conflicts between participants of a social movement and the general public. Previous studies Harel et al. (2020); Bar-Tal (2007) notice that ideological divergence and emotional confrontation are the micro-foundations of social conflicts in cyberspace. Conflicts exist in both intergroup and intragroup interactions, and outgroup pressure leads to further polarization inside a specific social group. Meanwhile, social media amplify the emotional factors that drive the conflicts. I follow this micro-interaction-based and emotion-driven approach to study the process of stigmatization of social movements in a dynamic process, discussing how the macro outcome of stigmatization emerges from micro-interactions between participants and non-participants of a social movement online.

I first go through the informal literature related to the conflicts in social movements, narrowing down the analytical scope of stigmatization to its micro mechanism—people's online interaction, and arguing that conflicts can manifest themselves through the device of emotional antagonism. Second, I discuss the limitation of existing formal models on their application to the question of interest. Then, using evolutionary game theory, I build a model to capture the dynamics of opinion evolution and conflict between two populations—the participants of a social movement and the general public. In the model, participants are either hostile or moderate towards the general public, while the outsiders of the movement are either haters or sympathizers. Conflicts happen inside the movement between hostile and moderate participants of a social movement; outside the movement, only moderate participants and sympathizers maintain peace. The assumption that a peaceful person loses more utility in conflicts than an emotionally hostile person drives all the implications of the model.

The model has several theoretical implications. First, hostile participants of a social movement and haters in the general public conflict with but also reproduce each other. Second, there exists a stigmatization threshold so that if there are enough hostile members of a social movement and haters in the general public, the evolutionary outcome is the stigmatization equilibrium in which all players choose the hostile/hater strategy. The larger the initial proportion of haters in the general public, the less hostile a social movement is needed to induce this stigmatization equilibrium. Third, the stigmatization of social movements is a self-fulfilling prophecy. In a specific situation, the social movement is not hostile at the beginning and is winning more sympathizers outside the movement, but enough level of initial stigmatization (haters in the general public) is creating more hostile participants of a social movement. If the latter effect is stronger than the former effect, then hostility toward the social movement still crosses the stigmatization threshold. In the situation above, severe stigmatization at the beginning produces a social movement that is antagonistic to the general public in the end. In others words, when the general public expect the movement to be extreme at the beginning, the movement will become extreme at the end.

2 Literature review

2.1 Conflicts, stigmatization, and emotions in social movements

Some social movements suffer severe stigmatization among the public (?Kim, 2018; Bavoleo and Chaure, 2020; Yang, 2014; Ka, 2021). Take the feminist movement worldwide as an example. South Korea's backlash against radical feminism became a solid voting base for conservative parties in the 2022 presidential election(Kim and Lee, 2022). In China, feminist expression online easily suffer from trolling, (Mao, 2020) and the antagonism between men and women has become a salient phenomenon in Chinese cyberspace [1]. In India, feminism is widely perceived as a western ideological invasion and hated among male anti-feminist Nicholas and Agius (2018); Rothermel (2020). Conflicts are inevitable between the thriving, progressive force and reactionary social structure (Pachter, 1974). The reactionary structure and its apparatuses stigmatize their challengers and have many names—racism, sexism, homophobia, and so on. On the individual level, stigmatization happens when some people hold false beliefs that a social movement and its participants are negative. Therefore, stigmatization is ideological instead of material. Scholars discuss who stigmatizes and why they stigmatize. The privileged group usually has limited empathy for the oppressed (DiAngelo, 2018). It thus holds negative perceptions of social movements that try to address social injustice and stigmatize them as"seeking privilege" and "creating oppression "(Dignam and Rohlinger, 2019; Van Valkenburgh, 2021).

[1] see https://www.bbc.com/zhongwen/simp/chinese-news-55571627

However, social movements also find their allies in the general public. Whether the privileged can be recognized as participants of a social movement has always been debated. For example, they De Beauvoir (2010) argued that men could not be called feminists because they lacked life experience as women. Putting aside the naming problem, apart from those who already identify themselves as participants of the movement, some sympathizers are generally pro-movement or open to its ideas. For instance, the support of African American men is essential in the practice of Black Feminism—empathy and support, if not "participation", were made possible by the unity legacy of shared participation in the civil rights movement (hooks, 2000).

Besides stigmatization and conflict outside the social movements, there are also intergroup conflicts between participants with different ideologies in a social movement. Harel et al. (2020) Shows that unaligned ideology in reciprocal interactions is an important foundation for conflicts to happen. Ideological divergence translates into hostile emotion and behavior, and intragroup ideological pressure further leads to opinion polarization and friction inside the socio-political group. Work Bar-Tal (2007) argues that in intractable social conflicts, people justify their own group and stigmatize other groups as "evil". Hence, ideological divergence causes conflict and stigmatization. In social movements mainly, conflict happens between the moderate and radical factions. The Baader-Meinhof (Aust, 2009) complex was one of the most extreme parts of the radical left in West Germany, and it finally broke up with the more moderate left-wing, resorting to terrorism to achieve its cause. The *Scum Manifesto*Solanas and Avital (2004) broadcasts the voice of radical feminism, and this legacy is appreciated and criticized by different feminist practices.

The discussion above indicates that first, there are several different strategies in the framework of a social movement—participants with varying ideologies in the movement, sympathizers in the general public, and enemies in the general public; second, ideological divergence is the micro foundation of conflicts and stigmatization in social movements. Then, what kind of ideologies divergence mainly cause conflict and stigmatization? Koo (2020) understands conflicts and stigmatization in the South Korean radical feminist movement as the result of trolling behavior—which is mainly emotion-driven and does not involve clear political agenda. Feminists in such forums like Megalia and Womad were trolled because they were not aggressive to biological men enough or did not agree with the tactic of "mirroring misogyny". Kim (2018) argues that the trolling and gender essentialism of such feminist practice is criticized as nonconstructive by its opposition. Ka (2021) also documents that rather than political blueprints, stigmatization and conflicts are driven by hostile sentiments. She also notices that even though there is a high diversity of ideologies in a social movement, people use reduced tags to label themselves as "radical" and "moderate" feminists, which is an emotional division of how hostile you are towards men and patriarchy rather than their political agendas. For instance, participants of the *6B4T* feminist movement in China and South Korea usually identify themselves as radical feminists due to their strong criticism of patriarchy. However, they only adopt individualistic resistance such as "no marriage" to resist without radical demand for social reform *de facto*. Hence, the conflicts in social movements may not necessarily be between radical and moderate groups but between people with *hostile* emotions and *less hostile* emotions towards the status quo. Similarly, people in the anti-movement forces also have a higher emotional hostility towards the movement, while self-identified sympathizers are less hostile.

I consider this emotion-driven approach valuable to help explain conflicts and stigmatization of social movements in the social media era. I then discuss how emotion and hostility are amplified by social media infrastructure.

2.2 Social movements in social media era

In the era of social media, some mechanisms make online communities more prone to interpersonal emotional hostility and conflict. First, social media reduces the coordination cost of social movements Van Laer and Van Aelst (2010); Kelly Garrett (2006), facilitating the spread of opinion and creating more chances for online interaction related to a particular social movement. For example, Crossley (2015) argues that social media not only helps the mobilization of feminism but also amplifies the voice of feminist influencers and helps information diffusion. However, social media also has the same effect on the anti-movement force, intensifying the conflict between the two forces. Second, information segregation and the logic of community feedback help social media amplify the impact of group polarization (Sunstein, 1999), making homophily and opinion polarization is a common feature on the internet (McPherson et al., 2001). This opinion polarization is more likely to happen when the topic is political (Barberá et al., 2015), and it naturally translates into sentiment polarization (Harel et al., 2020)—more extreme emotional status of people overall. Empirical evidence shows that negative emotion (Del Vicario et al., 2016) is infectious, and someone who receives negative feedback tends to send more of such feedback to others in the futureCheng et al. (2014). Therefore negative emotion among the population reinforces itself on social media platforms. Finally, the anonymous online environment reduces the cost of trolling and other violence and amplifies the voices of fight-pickers in real life, contributing to conflicts (Bor and Petersen, 2022).

This branch of literature conveys that social media amplify ideological and emotional polarization. It also supports the leverage of an emotional-driven approach to understanding political phenomena on the internet. These studies concentrate on technological aspects and platform interaction logic, emphasizing the interaction between people and the dynamic nature of opinions and emotions. Next, I will show models related to the dynamic evolution of opinions and discuss where my model can build on.

2.3 Related Models

Linking macro and micro level of analysis is an important task for social scientist(Coleman, 1994; Sawyer and Sawyer, 2005; Raub et al., 2011). It is usual to set up a formal model to discuss how macro outcomes emerge from micro-interactions. Literature on the stigmatization of social movements and social media suggest that the micro foundation of stigmatization and conflicts in social movements is ideological divergence—the divergence of people's opinion and attitudes. Accordingly, we need a model that can explain stigmatization by showing the process of opinion evolution, linking the macro outcome with the micro mechanism.

Previous models that involve opinion evolution incorporate one single population only and consider "persuasion" in multiple forms are the mechanism. In other words, people change their minds because of social feedback (Banisch and Olbrich, 2019; Bertotti and Delitala, 2008; Chen et al., 2020; Di Mare and Latora, 2007; Yuan et al., 2021). Some benchmark models try to find the condition of convergence or explain the formation of opinion polarization. In contrast, others extend them in different social networks, under uncertainty, or in a multi-dimensional environment. No model so far involves two divergent populations with different interaction rules.

Furthermore, no model considers the effect of censorship, which is a crucial and omnipresent feature in autocracy. Therefore, incorporating censorship into the model of opinion evolution can enrich the knowledge of the effect of censorship. Currently, students of authoritarian politics have discussed the mechanism of censorship as Bayesian persuasion (Gehlbach et al., 2021) and its role in preventing collective action (King et al., 2013) and discouraging opposition but also bringing about self-censoring, which reduces the information capacity of an autocrat. Nevertheless, no study has yet examined the effect of censorship on the evolution of social movements.

My model contributes in three aspects. First, I involve the dynamics of opinion evolution of *two* heterogeneous population, participants of the social movement, and outsiders in the general public, rather than a single population. This captures the reality that intergroup and intragroup interaction both exists (Harel et al., 2020), and allows me to examine the effect of outgroup pressure (emotional hostility) on the opinion evolution of the social movement. Second, my model provides another possible approach to opinion evolution that results from netizens' population replacement and culture inheritance in a given environment, complementary to the approach of persuasion via social feedback in the existing literature. This setting is a weaker assumption that allows people to be stubborn—opinions as strategies are assigned and unchangeable once the players enter the game, rather than assuming that they easily change their minds due to others' opinions. Third, I incorporate the effect of censorship and Echo Chamber into my model as two extensions, generating new theoretical implications.

2.4 Comments on previous literature

The stigmatization of social movements happens due to ideological divergence between the participants of social movements and the general public. The macro outcome of stigmatization emerges from ideological conflicts in the micro-interaction of people online, and this requires a model that is able to link both macro and micro levels. In the social media era, such ideological conflicts happen on digital platforms, and emotional hostility is amplified by the technological properties of social media. Hence, emotional hostility affects the stigmatization of social movements and online conflicts more in the social media era, which requires the academia to incorporate the emotion-driven approach in understanding the dynamic process of the stigmatization of social movements.

Several models of opinion dynamics relate to my research question in that they discuss how people's opinion change and how macro pattern is produced through this dynamic. However, they are not enough to settle my question. First, they only contain one single population, neglecting the reality that conflicts exist in both intergroup and intragroup. Second, they also do not involve censorship, which widely exists in authoritarian regimes and has a significant impact on the stigmatization of social movements. Third, previous models only consider the persuasion effect of feedback as the driving force of the opinion dynamics, ignoring alternative approaches such as population replacement under evolutionary pressure. To build on previous literature, I set up an evolutionary game theory model to capture the opinion dynamics of social movements online. My model contains two populations —the participants and outsiders in the general public, considering evolutionary fitness and population replacement as a new approach to frame opinion dynamics and discuss the effect of censorship and Echo Chamber as model extensions. The implications of my model help the understanding of my research questions—how the stigmatization of social movements emerges from micro-level ideological and emotional conflicts among people.

3 Method

This paper is the first one to leverage evolutionary game theory as a tool to analyze the opinion dynamics of social movements. Hence, a basic introduction to evolutionary game theory and its suitability for my analysis is needed.

In traditional non-cooperative games, scholars assume players to be rational and strategic (Osborne et al., 2004; McCarty and Meirowitz, 2007). However, the strong assumption of rationality is not always the case. Apolitical random events impact voters' behavior, such as the occurrence of shark attacks influencing election outcome (Achen and Bartels, 2012); ignorance is an important explanatory variable of why voters support trade protectionism (Rho and Tomz, 2017). These examples suggest that in discussing decentralized group behavior, it is more realistic to set up weaker assumptions about the rationality of players. In evolutionary game theory, players are considered to be non-strategic, so their strategies (genes) are assigned once they enter the game, and it never changes. Therefore, in a basic evolutionary game theory setup, a player's strategy is the same as her type. The proportions of different strategies in a certain population change according to natural selection pressure—different strategies have different levels of fitness (payoffs) as a function of the overall strategy profile (the proportions of all strategies of all populations), and the probability that a certain strategy leave "offspring" is positively correlated with their fitness. Hence, the fittest survive in the evolutionary game, and strategies with low payoffs decline.

As a modeling method initiated in biological studies, is it feasible for social scientists to apply evolutionary game theories? Weibull (1997) and Friedman (1998) both introduce the evolutionary game to social science and argue that evolutionary game theory is of great value for analyzing large-scale decentralized behavior of players (individuals, firms, and so on). Different from biology, the inheritance of strategies in social science is through norms, culture, and social networks. For the research question of this paper—how the stigmatization of social movements emerges from online interaction—the targeted population (netizens) is vast, and it is natural to consider them non-strategic. Empirical evidence shows that the behavior and opinion of new users of a specific social media platform are strongly decided by the pattern of current existing users (Cheng et al., 2014), which satisfies the "inheritance" approach of evolutionary games. Furthermore, the evolutionary game model assumes random matching between players, which is common in social media subgroups where everyone has access to interact with other members. It is easy to incorporate the effect of homophily to make the matching rule more realistic. Therefore, an evolutionary game is suitable to depict the dynamics of opinions in a specific online community. Readers should restrict the interpretation of my model to small to medium-size online communities where netizens intensively interact with each other.

The next section is a detailed introduction of the baseline model setup, including the basis of specific settings and the justification of the assumptions.

4 Model

There are two sub-populations in the game: participants of a social movement A and outsiders in the general public B. The whole population is normalized to 1, so that $A + B = 1$. Among the participants exists two strategies, hostile A_1 and moderate A_0 and among the general public exists two strategies, haters B_1 and sympathizers B_0. Table 1 shows the conflict-peace relations between the four strategies.

The two strategies of participants of a social movement is divided by their emotional attitudes—be friendly/hostile to whom. Hostile participants (A_1) consider moderate

Table 1: Conflict Relations Between The Four Strategies

	A_1 **Hostile**	A_0 **Moderate**	B_1 **Hater**	B_0 **Sympathizer**
A_1 **Hostile**	Peace	Conflict	Conflict	Conflict
A_0 **Moderate**	Conflict	Peace	Conflict	Peace
B_1 **Hater**	Conflict	Conflict		
B_0 **Sympathizer**	Conflict	Peace		

participants and all individuals outside the movement evil and only maintain peace with their kind. Examples of this strategy could be gender essentialist feminism discussed in Koo (2020). Readers should notice that this is an ideal type. Meanwhile, moderate participants (A_0) can maintain peace with their own kind and moderate outsiders but will conflict with any hostile individual. Among the general public, the sympathizers B_0 can interact peacefully with moderate participants of the social movement but not the hostile participants; the haters B_1 hate the movement and conflict with both types of participants. The alternative rightists make ideal examples of the "hater" strategy in online interactions (Van Valkenburgh, 2021; Dignam and Rohlinger, 2019). The level of stigmatization of a social movement is represented by the proportion of players in population B that adopt the hater strategy.

The initial proportions of all strategies are decided by factors not included in the model, such as culture, propaganda, or historical memory. The strategies of players when $T = 0$ refers to their initial expectation of how other players' strategies are, because it is before actual interactions.

The participants interact with both participants and outsiders, while outsiders only play with participants. This setup is justified that a participant may spend much time getting information from and chatting with both their peers in the movement (for alliance or opinion exchange) and outsiders (for propaganda and voicing for the movement). In contrast, a typical outsider in the general public spends limited time deciding whether to hate or sympathize with the movement based on their interaction with the participants. Outsiders do not spend much time discussing the movement with their peers (there are so many social movements, but so much work and so little time!), which means the game between outsiders is trivial.

The baseline model adopts random matching between players. The chance that a player meets a certain kind of strategy depends on the global proportion of that strategy. The definition of "interaction" in this study is general—it can be direct and dyadic communication between two players, one-way observation of information from another player, or small group discussion—and could be seen as an abstraction of online information flows.

Table 2: The Strategic Form of Fitness

	A_0	A_1	B_0	B_1
A_0	$1, 1$	$1-e, 1-m$	$1, 1$	$1-e, 1-m$
A_1	$1-m, 1-e$	$1, 1$	$1-m, 1-e$	$1-\gamma e, 1-\gamma e$

The fitness form (Table 2) of an evolutionary game is similar to the strategic form of payoffs in traditional non-cooperative games. The fitness is substantively ordinal, decided by several assumptions. First, people lose psychological utility from conflicts online. Second, in conflicts, people with moderate attitudes (A_0 and B_0) lose more utility than hostile people (A_1 and B_1), which is the critical assumption of this model. This assumption is inspired from the theory of emotional contagion Del Vicario et al. (2016)—the contagious nature of antagonism shows the supermodularity of antagonistic emotion so that it should be a dominant strategy when facing another player with hostile strategy. The second assumption means that $m < e \in (0, 1)$ and $\gamma \in (0, 1)$. Let $a_{\alpha,\beta}$

denotes the fitness when strategy α meets β. The fitness of a strategy is:

$$u_{A_i} = \sum A_j a_{A_i,A_j} + \sum B_j a_{A_i,B_j} \tag{1}$$

$$u_{B_i} = \sum A_j a_{B_i,A_j} \tag{2}$$

After clarifying the fitness form, we can calculate the replicator dynamics—the growth rate of a particular strategy in terms of the proportion in its own *sub-population* with respect to continuous time parameter t. The decision of signs of the replicator dynamics are simple: the proportions of those strategies with higher-than-average fitness in their own *sub-population* grow. The following equations specify the growth rate of a strategy:

$$\frac{\partial A_i}{\partial t} = A_i(u_{A_i} - \frac{A_0 u_{A_0} + A_1 u_{A_1}}{A_0 + A_1}) = A_i A_{-i}(u_{A_i} - u_{A_{-i}}) \tag{3}$$

$$\frac{\partial B_i}{\partial t} = B_i(u_{B_i} - \frac{B_0 u_{B_0} + B_1 u_{B_1}}{B_0 + B_1}) = B_i B_{-i}(u_{B_i} - u_{B_{-i}}) \tag{4}$$

The sequence of the game is:

1. Nature chooses the initial $\boldsymbol{x}_{t_0}$=(A_{1t_0}, B_{1t_0}) by a generic distribution F, so that the proportions of all four strategies are decided.

2. All players in the whole population interaction with each other in time t according to a certain matching rule (pure random matching in the baseline model).

3. The fitness of the strategies is realized by aggregating the fitness of players that adopt it.

4. The proportion of the strategies grow/decline according to the replicator dynamics

until $\boldsymbol{x}$ converges to any evolutionary stable state (ESS).

5 Analysis

By looking at the fitness form, we infer that this game is similar to a coordination game where players across populations want to match their types. In other words, if new entering players in time t anticipate other players more likely to be moderate (A_0 or B_0), they are more likely to inherit the moderate/sympathizer strategy, otherwise they are more likely to inherit the hostile/hater strategy. There exists three Nash equilibria in a basic coordination game, but the unique interior solution is not an ESS because it does not have a uniform barrier of invasion (Weibull, 1997) so that any external shock at the interior Nash equilibrium will lead to convergence to the other two ESSs. Therefore, there are two ESSs in the game.

Proposition 1 (Evolutionary Stable States) *There exists two evolutionary stable state. First, the peace equilibrium $\boldsymbol{x}^*_P = \{A_1 = 0,\ B_1 = 0\}$ in which all players are moderate participants or sympathizers of the movement; second, the stigmatization equilibrium $\boldsymbol{x}^*_S = \{A_1 = 1,\ B_1 = 1\}$ in which all players are hostile participants or haters of the movement.*

The proof of which is trivial. Proposition 1 shows that in any ESS, both participants and outsiders match their type. The first ESS in which all players adopt the moderate or sympathizer strategies is the *peace equilibrium* and there is no conflict and stigmatization, while the second ESS in which all players adopt hostile or hater strategies is the *stigmatization equilibrium*, and there are conflicts and stigmatization between the movement and the general public. The result is not intuitive since it is common that opinion divergence exists in the movement and the general public empirically. However, readers should notice that, first, the interaction environment in the model is where all players can interact with each other, hence the equilibrium result is restricted to certain online communities, forums, or platforms, rather than the internet as a whole; second, the equilibrium can be interpreted as the tendency and direction of evolution according with respect to time t, therefore what we observed in real life can be still in-progress of the evolution.

Now I analysis the dynamics of evolution. Before the global mixed strategy $\boldsymbol{x}$ converge to any ESS, the proportion of strategies change according to the signs of replicator dynamics. Because there are only two strategies in a certain sub-population, if the proportion of one strategy grows, the other one must decline. Therefore, the conditions for the proportions of moderate participants A_0 and sympathizers B_0 to grow, so that the proportions of hostile players A_1 and B_1 decline, are given by:

$$u_{A_0} > u_{A_1} \tag{5}$$

$$u_{B_0} > u_{B_1} \tag{6}$$

which solve four domains divided by two curves, $C1$ and $C2$ so that:

$$C1: \quad A_1 = \frac{m}{e + m - e\gamma} \tag{7}$$

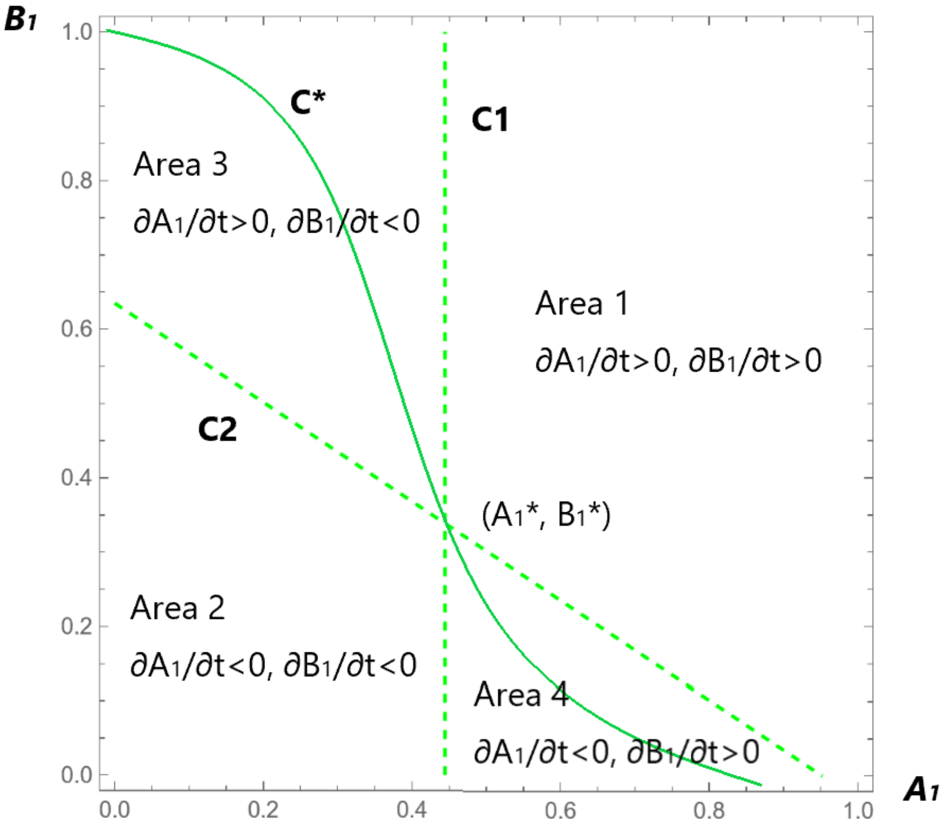
$$C2: \quad B_1 = -\frac{(ep + mp)A_1}{(1-p)(e + m - e\gamma)} + \frac{m}{(1-p)(e + m - e\gamma)} \tag{8}$$

and the four domains are shown in Figure 1. In figure 1, point $(A_1^*,\ B_1^*)$ is the interior Nash equilibrium of the game, $C1$ satisfies $u_{B_0} = u_{B_1}$ and $C2$ satisfies $u_{A_0} = u_{A_1}$. Denote the proportions of strategies in time t as $\boldsymbol{x}_t$. If $\boldsymbol{x}_t$ is in Area 1, A_1 and B_1 both grow, and the game converges to the stigmatization equilibrium $\boldsymbol{x}_S^*$, else if $\boldsymbol{x}_t$ is in Area 2, A_1 and B_1 both decline, and the game converges to the peace equilibrium $\boldsymbol{x}_P^*$. This result is intuitive and reflect the situations in reality that when both participants and outsiders of a social movement is hostile (moderate) enough, the relationship between the movement and the general public is high-conflict (peaceful). The property of replicator dynamics in Area 1 and 2 is formalized as Lemma 1, which is important for analysis of Area 3 and 4.

Lemma 1 (Replicator Dynamics in Area 1 and 2) *In any time t, if $\boldsymbol{x}_t = \{A_{1t}, B_{1t}\}$ satisfies that $A_{1t} > \frac{m}{e+m-e\gamma}$ and $B_{1t} > -\frac{(ep+mp)A_{1t}}{(1-p)(e+m-e\gamma)} + \frac{m}{(1-p)(e+m-e\gamma)}$, the evolutionary outcome converges to $\boldsymbol{x}_S^*$ when $t \to +\infty$; if $\boldsymbol{x}_t = \{A_{1t}, B_{1t}\}$ satisfies that $A_{1t} < \frac{m}{e+m-e\gamma}$ and $B_{1t} < -\frac{(ep+mp)A_{1t}}{(1-p)(e+m-e\gamma)} + \frac{m}{(1-p)(e+m-e\gamma)}$, the evolutionary outcome converges to $\boldsymbol{x}_P^*$ when $t \to +\infty$.*

It is more interesting to discuss dynamics in Area 3 and 4, In these two areas, the A_1 and B_1 evolve in the opposite direction. In Area 3, there are too many haters B_1 in the general public so that the proportion of hostile participants grows as retaliation, while at the same time, because there are enough moderate participants to interact peacefully with the general public, the proportion of sympathizers of the movement also grows. Situation in Area 4 mirrors the dynamics in Area 3 so that the social movement is becoming moderate while the proportion of its haters in the general public grows. Clearly, the evolutionary outcome in Area 3 and 4—whether it converges to the peace equilibrium $\boldsymbol{x}_P^*$ or the stigmatization equilibrium $\boldsymbol{x}_S^*$—depends on the relative speed at which two sub-populations evolve in opposite directions. Take dynamics in Area 3 for example, if the speed at which the movement becomes more hostile overwhelms the trend of the general public to be more friendly to the movement, $\boldsymbol{x}_t$ will reach Area 1 at some t, and according to Lemma 1, the game ends with $\boldsymbol{x}_S^*$.

Figure 1: Absorption Domain of ESSs



To complete the discussion of dynamics in Area 3 and 4, I now formally derive the absorption domains of the two ESSs, which is equivalent to derive the solution of the initial value problem for an ordinary differential equation system $f(t, \boldsymbol{x})$.

$$f(t, \boldsymbol{x}_{t_0}) = \begin{cases} \dfrac{\partial A_1}{\partial t} \\ \dfrac{\partial B_1}{\partial t} \\ A_1(0) = A_{1t_0} \\ B_1(0) = B_{1t_0} \end{cases} \tag{9}$$

Particularly, the initial value problem $f(t, (A_1^*, B_1^*))$ solves a unique curve C^* that announces the absorption domains of the two ESSs. Therefore, C^* is the *Stigmatization Threshold.*

Proposition 2 (Stigmatization Threshold) *There exists a unique curve C^* that goes through the interior Nash equilibrium (A_1^*, B_1^*). Any $\boldsymbol{x}_{t_0}$ below C^* leads to $\boldsymbol{x}_P^*$ and any $\boldsymbol{x}_{t_0}$ above C^* leads to $\boldsymbol{x}_P^*$.*

The proof of Proposition 2 is in the appendix. The baseline model generates three theoretical implications. First, the dynamic of large enough hostility and stigmatization between the movement and the general public is a process of reinforcing each other. Second, the dynamics in Area 3 shows that stigmatization of a social movement can be a self-fulfilling prophecy when the proportion of haters are large enough. Haters in the general public stigmatize and attack all participants of the movement, which in turns creates more hostile participants in the movement and if this "hostility as retaliation" effect is strong enough, everyone becomes hostile to each other in the end. Third, the shape of the stigmatization threshold C^* indicates that if one side of the game, participants or outsiders, is more hostile/hating, it lowers the threshold of sufficient hostility of the other side to predict a stigmatization equilibrium.

6 Extensions

6.1 Echo Chamber

Homophily is a salient phenomenon in today's society, especially in cyberspace (Boutyline and Willer, 2017; Barberá et al., 2015; McPherson et al., 2001; Fiore and Donath, 2005; Sánchez et al., 2016). People tend to interact with other people of similar ideas, creating an Echo Chamber that exacerbate the homogeneity of opinions locally. Echo Chamber effect is strongly correlated to opinion polarization. Intuitively, the Echo Chamber effect seems relate itself with more stigmatization and conflict online. But is it the case in this model?

The baseline model predict two ESSs, which present convergence instead of polarization within a single population, however, the two equilibrium are different in terms of the intragroup convergence-polarization outcome. In the peace equilibrium, both participants and outsiders are inclusive to each other without conflict, thus meaning convergence between the two sub-populations; in the stigmatization equilibrium, though both population are internally convergent, they attack each other in intragroup interaction, thus is more close to polarization. To be clear, in this model, the stigmatization equilibrium is equivalent to polarization. In this equilibrium, players have the right beliefs about other players in the end, but the self-fulfilling nature of the bad equilibrium indicates that there can be misunderstandings at the beginning of the game.

To incorporate the Echo Chamber effect, it is convenient to set a fixed probability h_i that the strategy A_i meet itself in time t, and probability $1 - h_i$ that A_i is randomly matching with all other strategies just like in the baseline model. The fitness of B is unchanged. The new fitness function for A is:

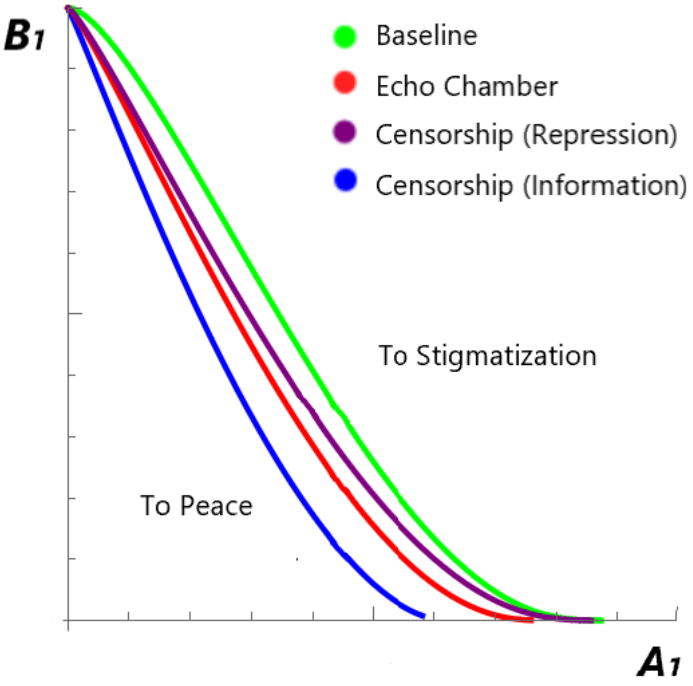
$$u_{hA_i} = h_i + (1 - h_i)(\sum A_j a_{A_i,A_j} + \sum B_j a_{A_i,B_j}) \tag{10}$$

which is an affine transformation of the fitness function in baseline. If $h_0 = h_1 = h$, the sign of $\frac{\partial A_i}{\partial t}$ is decided by $u_{hA_i} - u_{hA_-i} = (1 - h)(u_{A_i} - u_{A_-i})$, which means that when both strategies have the same level of homophily, the Echo Chamber effect only changes the velocity of replicator dynamics, making the elimination of one strategy slower, but does not change the evolutionary result. Only when $h_1 > h_0$, which means hostile participants A_1 are more like "birds of a feather" than moderate participants A_0, will the Echo Chamber effect contributes to stigmatization.

Proposition 3 (The Effect of Echo Chamber) *There is a unique stigmatization threshold C_h^* that goes through the interior Nash equilibrium $\{A_{1h}^*, B_{1h}^*\}$. C_h^* is below C^* if and only if $h_1 > h_0$. Echo Chamber enlarges the strategy space that induces $\boldsymbol{x}_S^*$ if and only if the homophily effect of A_1 is stronger than A_0.*

The proof of Proposition 3 is in the appendix. When the Echo Chamber effect among the hostile participants is stronger than it among the moderate participants, the strategy space that induces the peace equilibrium is smaller than in the baseline model because the stigmatization threshold C_h^* in the Echo Chamber model is strictly below the threshold in the baseline model (see figure 2). Hence, surprisingly, in our model, homophily does not necessarily help stigmatization, but different level of homophily between different strategies does. The intuition is that while Echo Chamber among the hostile participants secures more psychological payoff for new entering users to be hostile, it has the same effect of encouraging users to be moderate. When people talk about Echo Chamber and polarization, they assume if people have more exposure to diverse information, there will be less polarization, so people's opinion is a balance between different source of information. However, in this model, I show that the interaction with outside information does not necessarily make a person similar to them, but instead, when the interaction is not satisfactory, it strengthens the existing biases. The counter-intuitive result arises because if moderate people have more "protection" from emotionally hostile content, they are more like to stick to their strategy. This is the case when social media is not founded and people with extreme stances are hard to find their allies. Therefore, Echo Chamber effect on social media helps more for extreme opinions rather than moderate opinions because such Echo Chamber does not widely existed before social media era.

Figure 2: The Effect of Echo Chamber and Censorship on Stigmatization



6.2 Censorship

What is censorship? Usually censorship is considered as information manipulation that selectively delete information or even prevent the generation of information. Yet in the spirit of Adena et al. (2015), censorship could also be a form of repression that warning someone that her opinion is not appreciated. I incorporate both effects of censorship by setting up a probability b that interaction with a certain kind of strategy is blocked (information manipulation) and a dead weight cost c on the fitness of strategies (repression).

Similar to the previous extension, we know that if the effect of censorship on the fitness is symmetric for both A_1 and A_0, then the replicator dynamics is unchanged. Therefore, for parsimony and clarity, I only add the treatment on the "more censored" strategy.

In autocracy, censorship mainly targets posts with collective action potential (King et al., 2013) and alternatives of the official ideology (Shue, 1990) which is generated more by the moderate participants of a social movement, rather than hatred speech, trolling, or other forms of offense and language violence generated by hostile participants. Take the *Douban* forum, the headquarter of Chinese feminist movement, as an example, censorship targets at *Douban* groups that systematically criticize patriarchy and voice clear political agenda, rather than the groups that express hatred towards biological men or troll their enemies[2], which makes moderate participants instead of hostile participants more vulnerable to censorship because they are emotionally moderate but politically radical while participants with hostile emotions are the opposite.

First consider the effect of censorship as repression:

$$u_{cA_0} = \sum A_j a_{A_0,A_j} + \sum B_j a_{A_0,B_j} - c \tag{11}$$

$$u_{cA_1} = u_{A_1} \tag{12}$$

and then as information manipulation:

$$u_{bA_i} = A_1 a_{A_i,A_1} + (1-b) A_0 a_{A_i,A_0} + \sum B_j a_{A_0,B_j} \tag{13}$$

$$u_{bB_i} = A_1 a_{B_i,A_1} + (1-b) A_0 a_{B_i,A_0} \tag{14}$$

Both setups reduce the fitness of moderate participants. In time t, the fitness of u_{cA_0} in the

[2] see working paper by Hongding Zhu and Jose E. Trinidad, available on https://docs.google.com/document/d/1stlpfjHY8wUfXEM0Q_anE9LcMmB6AEQJRIzd4qYphbY/edit?usp=sharing

repression extension is smaller than in the baseline model and accordingly the replicator dynamics of A_1 is larger than in the baseline model. So censorship as repression is in favor of stigmatization. Intuitively, the fitness of moderate strategy is reduced by censorship, and more participants of the social movements choose to be hostile than in the baseline model, which in turn results in more stigmatization (more haters) of the movement.

Proposition 4 (The Effect of Censorship) *For any $b, c > 0$, there is a unique stigmatization threshold C_c^* that goes through the interior Nash equilibrium $\{A_{1c}^*, B_{1c}^*\}$. C_c^* is below C^*. Censorship enlarge the strategy space that induces $\boldsymbol{x}_S^*$.*

The proof of Proposition 4 is trivial because it uses exactly the same logic and steps as Proposition 3. When participants of a social movement suffer extra cost for articulating their ideas and agenda in peaceful language, they gradually turn to trolling and emotional hostility. This story is the case in the Chinese cyberfeminist movement. Feng (2017); Shen (2021) shows that censorship creates a narrower space for feminists in China to articulate themselves while stories of “how censorship drives me to become an extreme feminist” are common on the *Douban* forum and Twitter.

7 Conclusion

I set up a two-population evolutionary game to model the opinion evolution and dynamics of stigmatization of a social movement in cyberspace. The interaction is among participants of the movement and between the participants and the general public.

Based on the critical assumption of “hostility as retaliation”, hostile participants and haters in the general public tend to reproduce each other. If the initial proportion of hostile participants of the movement exceeds the stigmatization threshold, the evolutionary outcome is the stigmatization equilibrium, where all participants are hostile to the public, and all outsiders are haters of the movement. The stigmatization threshold is pinned down by the proportion of haters in the general public so that the higher the initial stigmatization, the lower the threshold to induce the stigmatization equilibrium. I also discuss the typical situation in which stigmatization of a social movement is a self-fulfilling prophecy. In this situation, the movement itself is moderate at the beginning; however, severe enough social stigmatization (haters in the general public) will drive the movement to become more and more hostile, and finally, the conflict between the public and the movement is inevitable.

The effect of the Echo Chamber and censorship on the stigmatization of social movements are discussed in the extensions. The analysis shows that Echo Chamber only helps stigmatization and polarization when hostile participants in a social movement have a larger tendency toward homophily. Moreover, censorship in autocracy usually targets collective action potential and alternative ideologies other than the official one instead of trolling and language violence, which means that moderate participants, instead of hostile participants, are more vulnerable to censorship. Consequently, censorship exacerbates the stigmatization of social movements in autocracies, no matter whether it is interpreted as repression or information manipulation.

These results contribute to the theoretical discussion from several aspects. First, it analyzes the process of stigmatization of social movements through a dynamic approach, which complements the static analysis in the literature. The mechanism of stigmatization as a self-fulfilling prophecy helps explain how an initially moderate and peaceful social movement can descend into bitter conflict with its opponents in society. Second, in addition to complex ideological and structural factors, stigmatization of social movements can also arise from micro-emotional conflicts in online interactions, which is a crucial feature in the social media era and is supported by empirical observations (Koo, 2020; Ka, 2021). Third, different from previous studies (Harel et al., 2020) that claim outgroup pressure intensifies opinion polarization inside a social group, my model suggests that in a micro online community, stigmatization from the public induces only intragroup polarization and creates intergroup convergence—members of both groups choose internally consistent but externally hostile strategies. Hence, we should reconsider the does the previous claim hold only at a macro level. And finally, the logic of evolutionary games suggests that in studying the opinion dynamics, population replacement and ideological inheritance from old users to new users is a complementary mechanism other than persuasion.

Readers should also be aware of the limitations of this study. First, the assumption of random matching in the model restricts its analytical level to a micro online community where the discourse of a social movement circulates rather than the global dynamics on the internet. Second, the four strategies in the game are ideal types, and this typology captures only limited aspects of individual opinions in social movements. Third, the theoretical implications of the model require further empirical evidence to prove its credibility.

References

Christopher H Achen and Larry M Bartels. Blind retrospection: Why shark attacks are bad for democracy. *Center for the Study of Democratic Institutions, Vanderbilt University. Working Paper*, 2012. 10

Maja Adena, Ruben Enikolopov, Maria Petrova, Veronica Santarosa, and Ekaterina Zhuravskaya. Radio and the rise of the nazis in prewar germany. *The Quarterly Journal of Economics*, 130(4):1885–1939, 2015. 22

Stefan Aust. *Baader-Meinhof: The inside story of the RAF.* Oxford University Press, USA, 2009. 4

Sven Banisch and Eckehard Olbrich. Opinion polarization by learning from social feedback. *The Journal of Mathematical Sociology*, 43(2):76–103, 2019. Publisher: Taylor & Francis. 7

Daniel Bar-Tal. Sociopsychological foundations of intractable conflicts. *American Behavioral Scientist*, 50(11):1430–1453, 2007. 1, 4

Pablo Barberá, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10):1531–1542, 2015. Publisher: Sage Publications Sage CA: Los Angeles, CA. 6, 19

Bárbara Inés Bavoleo and Desireé Chaure. Korean cyberfeminism: emergence, characteristics and results. *Anuario en Relaciones Internacionales del IRI*, 2020, 2020. 3

Maria Letizia Bertotti and Marcello Delitala. On a discrete generalized kinetic approach for modelling persuader's influence in opinion formation processes. *Mathematical and computer modelling*, 48(7-8):1107–1121, 2008. Publisher: Elsevier. 7

Alexander Bor and Michael Bang Petersen. The psychology of online political hostility: A comprehensive, cross-national test of the mismatch hypothesis. *American political science review*, 116(1):1–18, 2022. 6

Andrei Boutyline and Robb Willer. The social structure of political echo chambers:

Variation in ideological homophily in online networks. *Political psychology*, 38(3):551–569, 2017. Publisher: Wiley Online Library. 19

Tinggui Chen, Qianqian Li, Peihua Fu, Jianjun Yang, Chonghuan Xu, Guodong Cong, and Gongfa Li. Public opinion polarization by individual revenue from the social preference theory. *International journal of environmental research and public health*, 17(3):946, 2020. Publisher: Multidisciplinary Digital Publishing Institute. 7

Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. How community feedback shapes user behavior. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014. 6, 11

James S Coleman. *Foundations of social theory*. Harvard university press, 1994. 7

Alison Dahl Crossley. Facebook feminism: Social media, blogs, and new technologies of contemporary us feminism. *Mobilization: An International Quarterly*, 20(2):253–268, 2015. 6

Simone De Beauvoir. *The second sex*. Knopf, 2010. 4

Michela Del Vicario, Gianna Vivaldo, Alessandro Bessi, Fabiana Zollo, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Echo chambers: Emotional contagion and group polarization on facebook. *Scientific reports*, 6(1):1–12, 2016. 6, 13

Alessandro Di Mare and Vito Latora. Opinion formation models based on game theory. *International Journal of Modern Physics C*, 18(09):1377–1395, 2007. Publisher: World Scientific. 7

Robin DiAngelo. *White fragility: Why it's so hard for white people to talk about racism*. Beacon Press, 2018. 3

Pierce Alexander Dignam and Deana A Rohlinger. Misogynistic men online: How the red pill helped elect trump. *Signs: Journal of Women in Culture and Society*, 44(3): 589–612, 2019. 3, 12

Feng. Hard times for feminists in China. *SupChina*, 2017. URL https://supchina.com/2017/03/08/hard-times-feminists-china/. 1, 24

Andrew T Fiore and Judith S Donath. Homophily in online dating: when do you like someone like yourself? In *CHI'05 extended abstracts on Human factors in computing systems*, pages 1371–1374, 2005. 19

Daniel Friedman. On economic applications of evolutionary game theory. *Journal of evolutionary economics*, 8(1):15–43, 1998. 10

Scott Gehlbach, Zhaotian Luo, Anton Shirikov, and Dmitriy Vorobyev. A Model of Censorship, Propaganda, and Repression. 2021. 8

Noriko Hara and Bi-Yun Huang. Online social movements. 2011. 1

Tal Orian Harel, Ifat Maoz, and Eran Halperin. A conflict within a conflict: intragroup ideological polarization and intergroup intractable conflict. *Current Opinion in Behavioral Sciences*, 34:52–57, 2020. 1, 4, 6, 8, 26

bell hooks. *Feminist theory: From margin to center*. Pluto Press, 2000. 4

Fei Ka. douban nvquan, jiafeng zhong xingcheng de "zhongguo tese nvquan". *Initium Media*, March 2021. URL https://theinitium.com/article/20210308-opinion-china-douban-feminism-awakens-or-failure/?utm_source=twitter&utm_medium=twitter&utm_campaign=twpost. 3, 5, 26

R Kelly Garrett. Protest in an information society: A review of literature on social movements and new icts. *Information, communication & society*, 9(02):202–224, 2006. 6

Bo-Myung Kim. Late modern misogyny and feminist politics: The case of ilbe, megalia, and womad. *Journal of Korean Women's Studies*, 34(1):1–31, 2018. 3, 5

Hannah June Kim and Chungjae Lee. The 2022 south korean presidential election and the gender divide among the youth. *Pacific Affairs*, 95(2):285–308, 2022. 3

Gary King, Jennifer Pan, and Margaret E Roberts. How censorship in china allows government criticism but silences collective expression. *American political science Review*, 107(2):326–343, 2013. 8, 22

JiHae Koo. South korean cyberfeminism and trolling: The limitation of online feminist community womad as counterpublic. *Feminist Media Studies*, 20(6):831–846, 2020. 5, 12, 26

Chengting Mao. Feminist activism via social media in china. *Asian Journal of Women's Studies*, 26(2):245–258, 2020. 3

Nolan McCarty and Adam Meirowitz. *Political game theory: an introduction*. Cambridge University Press, 2007. 10

Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001. Publisher: Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA. 6, 19

Lucy Nicholas and Christine Agius. # notallmen,# menenism, manospheres and unsafe spaces: Overt and subtle masculinism in anti-"pc" discourse. In *The persistence of global masculinism*, pages 31–59. Springer, 2018. 3

Martin J Osborne et al. *An introduction to game theory*, volume 3. Oxford university press New York, 2004. 10

Henry M. Pachter. The idea of progress in marxism. *Social Research*, 41(1):136–161, 1974. ISSN 0037783X. 3

Werner Raub, Vincent Buskens, and Marcel ALM Van Assen. Micro-macro links and microfoundations in sociology. *The Journal of Mathematical Sociology*, 35(1-3):1–25, 2011. Publisher: Taylor & Francis. 7

Sungmin Rho and Michael Tomz. Why don't trade preferences reflect economic self-interest? *International Organization*, 71(S1):S85–S108, 2017. 10

Ann-Kathrin Rothermel. Global–local dynamics in anti-feminist discourses: An analysis of indian, russian and us online communities. *International Affairs*, 96(5):1367–1385, 2020. 3

R Keith Sawyer and Robert Keith Sawyer Sawyer. *Social emergence: Societies as complex systems*. Cambridge University Press, 2005. 7

Du Shen. kongjian jiya yu jixing shengzhang: zhongguo nvquan de neiyouwaikun. *Initium Media*, April 2021. URL https://theinitium.com/article/20210415-opinion-china-feminisim-statism/. 24

Vivienne Shue. *The reach of the state: sketches of the Chinese body politic.* Stanford University Press, 1990. 23

Valerie Solanas and Ronell Avital. *SCUM manifesto.* Verso, 2004. 4

Cass R Sunstein. The law of group polarization. *University of Chicago Law School, John M. Olin Law & Economics Working Paper*, (91), 1999. 6

Daniel López Sánchez, Jorge Revuelta, Fernando De la Prieta, Ana B Gil-González, and Cach Dang. Twitter user clustering based on their preferences and the Louvain algorithm. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*, pages 349–356. Springer, 2016. 19

Jia Tan. Digital masquerading: Feminist media activism in China. *Crime, Media, Culture*, 13(2):171–186, 2017. 1

Jeroen Van Laer and Peter Van Aelst. Internet and social movement action repertoires: Opportunities and limitations. *Information, Communication & Society*, 13(8):1146–1171, 2010. 6

Shawn P Van Valkenburgh. Digesting the red pill: Masculinity and neoliberalism in the manosphere. *Men and Masculinities*, 24(1):84–103, 2021. 3, 12

Melissa A Wall. Social movements and email: expressions of online identity in the globalization protests. *New media & society*, 9(2):258–277, 2007. 1

Jörgen W Weibull. *Evolutionary game theory.* MIT press, 1997. 10, 15

Yuke Yang. Jijin de nvquan biaoqian: Nvquan zhuyi ruhe zai meijie pingtai bei wuminghua. *Journalism and Communication*, 2014:94, 2014. 3

Jiangjun Yuan, Jiawen Shi, Jie Wang, and Weinan Liu. Modelling network public opinion polarization based on SIR model considering dynamic network structure. *Alexandria Engineering Journal*, 2021. Publisher: Elsevier. 7

Appendix

Proof of Proposition 2

Because the first-order ODEs $f(t, \boldsymbol{x})$ are polynomials about $\boldsymbol{x}$, they satisfy the Lipschitz condition and have one unique solution $\{A_1(t),\ B_1(t)\}$ according to Picard-Lindelöf Theorem. Which means any initial value problem has a unique evolutionary track—a unique curve that goes through it. Therefore, there exists a unique curve C^* in Figure 1 that goes through the interior Nash equilibrium $(A_1^*,\ B_1^*)$ and any $\boldsymbol{x}_{t_0}$ not on C^* never cross C^* during the replicator dynamics. Hence, any $\boldsymbol{x}_{t_0}$ below C^* must arrive at Area 2 at some t and any $\boldsymbol{x}_{t_0}$ above C^* must arrive at Area 1 at some t. And according to Lemma 1, the evolutionary outcome after arriving at Area 1 or 2 is clear. ■

Proof of Proposition 3

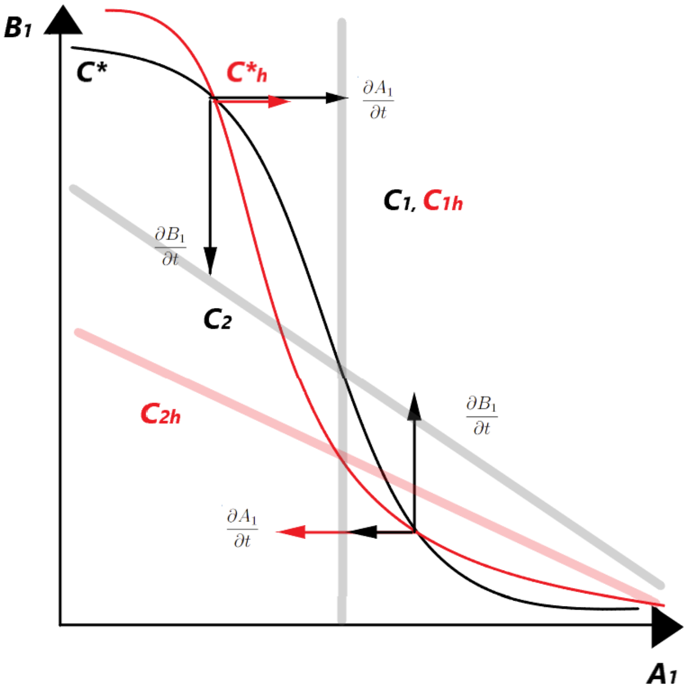
Denote the two curves which is the solution of $u_{hB_0} = u_{hB_1}$ and $u_{hA_0} = u_{hA_1}$ as $C1_h$ and $C2_h$, it solves:

$$C1_h: \quad A_1 = \frac{m}{e + m - e\gamma} \tag{15}$$

$$C2_h: \quad B_1 = \frac{(1-h)m}{(1-p)[e + (1-h)m - (1-h)e\gamma]} - \frac{p[e + (1-h)m]A_1}{(1-p)[e + (1-h)m - (1-h)e\gamma]} \tag{16}$$

in which $C1_h$ remains the same as $C1$ while $C2_h$ is strictly below $C2$ when $h_1 > h_0$. Therefore, the interior Nash equilibrium (A_{1h}^*, B_{1h}^*) is strictly below (A_1^*, B_1^*) of the baseline model. Accordingly, there is a unique stigmatization threshold C_h^* that goes through (A_1^*, B_1^*). Suppose C^*h crosses C^* at some point, then at the intersection point $(\tilde{A}_1, \tilde{B}_1)$, the replicator dynamics (growth rate) of B_1 is the same in the baseline and the Echo Chamber model, while the replicator dynamics of A_1 must be smaller than in the baseline model in the two situation in Figure 3. However, in the setup we know that on any point (A_1, B_1), the replicator dynamic of A_1 is larger than in the baseline model since $u_{hA_1} > u_{A_1}$ on any point (A_1, B_1). There are contradictions. So the intersection situations in Figure 3 does not exist, and accordingly C_h^* is always below C^*. ■

Figure 3: Suppose the Intersection Exists



B1
C*
C*h
$\frac{\partial A_1}{\partial t}$
C1, C1h
$\frac{\partial B_1}{\partial t}$
C2
C2h
$\frac{\partial B_1}{\partial t}$
$\frac{\partial A_1}{\partial t}$
A1