Discussion Paper Series
No. 2202
February, 2024

On the Origin of Polarization

By John Duffy, Seung Han Yoo

The Institute of Economic Research – Korea University
Anam-dong, Sungbuk-ku, Seoul, 136-701, South Korea, Tel: (82-2) 3290-1632, Fax: (82-2) 928-4948

On the Origin of Polarization*

John Duffy† Seung Han Yoo‡

February 2024

Abstract

We provide a dynamic, spatial model of group sorting and polarization based on observable group identity alone. Agents, who belong to one of two perfectly observable groups, are randomly matched to play an investment game in youth and old age, and make a location decision in between. Each agent's ability is private information and the true distribution of each group's of abilities (types) is uncertain, so agents have to form beliefs about these distributions in making both investment and location decisions. Further, agents have no preferences or special facilities for interacting with members of their own group. We characterize the equilibrium for this endogenous, dynamic spatial location model which involves the formation of higher order perceptions and we show that under certain conditions, a limiting outcome of the dynamical system is that the society becomes completely polarized with members of each group rationally choosing to congregate in distinct locations.

Keywords: matching, private monitoring, sorting, polarization, group bias, homophily, Bayesian learning.
JEL Classification Numbers: C72, C73, D83.

---

*We thank the participants at various conferences for comments on previous versions. The usual disclaimer applies. Yoo acknowledges that this research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. RS- 2023-00243709).

†Department of Economics, University of California, Irvine, California, 92697 (e-mail: duffy@uci.edu).

‡Department of Economics, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul, Republic of Korea, 02841 (e-mail: shyoo@korea.ac.kr).

1 Introduction

Individuals frequently choose to engage with others who are similar to themselves. This tendency toward *homophily* can encompass many dimensions such as race, ethnicity, income, culture, religious beliefs, educational attainment, and politics, and can result in the sorting of individuals both spatially, *e.g.*, into distinct homogeneous communities, or virtually, *e.g.*, by the sources of news they consume.[1] Current explanations for this phenomenon typically revolve around either personal preferences or communication costs. However, even in the absence of differences in such preferences or costs, *perceptions of others* may also play a crucial role. In this paper, we develop a dynamic spatial model showing how perfect sorting of group members to different locations, or "polarization," can be *perceptions-based*. One can think of our model as providing a kind of "statistical discrimination" rationalization for polarization that is distinct from the more common "taste-based discrimination" approach to polarization originating in the seminal work of Tiebout (1956) and Schelling (1971).

We model the *origin* of such sorting, leading to polarization, starting from seemingly innocuous initial conditions. Specifically, we consider a model where all individuals are born as members of one of two groups labeled Blue and Red. Group membership is perfectly identifiable, but in all other dimensions, including individual abilities (types) or the distribution of abilities by group, individuals are completely indistinguishable from one another. Importantly, neither group has any explicit preference for interacting with members of its own group or the other group, nor is there any cost difference in interactions within or between groups. Further, members of both groups are initially dispersed between two possible locations, East and West.[2] Beginning from such seemingly inconsequential initial conditions, we seek to understand the sorting of individuals into two perfectly polarized groups of all Red and all Blue, with each group occupying a single location, either East or West, based on perceptions alone and under rational belief updating.[3]

As with all origin stories, we need a plot device that does not strain credulity. The mechanism we employ involves *distributional uncertainty and private monitoring* together with Bayesian belief updating. Specifically, we consider the case where each group's types are drawn from one of just two possible distributions but the true distribution characterizing types for each group is unknown. While group membership is perfectly observable, a player only knows his own history of play with others, which can also be differentiated by group identity; that is, we assume private monitoring. For tractability reasons, we consider a setup

[1]This phenomenon has been well-documented by political scientists, *e.g.*, Huckfeldt and Sprague (1995), sociologists, *e.g.*, McPherson, Smith-Lovin and Cook (2001), and economists, *e.g.*, Currarini, Jackson and Pin (2010) and Goeree et al. (2010).

[2]In the model, we consider a finite number of locations, and show that in equilibrium, there are only two sets of locations where players will choose to locate themselves.

[3]We are not specific regarding the dimension on which players become polarized; it could be anything including politics, language, religion or race, among many possibilities.

where players live for just two periods. In each period, they can interact with members of either group but only with those of their own generation (or age). Importantly, young agents are born with unbiased beliefs; they think that both groups are equally likely to have the same type distributions and they update these beliefs based on their history of play. Young agents play the investment game in the location chosen by their parent and then decide whether to remain in that same location or move to the other location to play the same game again when they are old. The young agent's location choice depends on their history of play. While agents live for just two periods and have only one-period payoff relevant histories, the proportions of the group members in the two locations at each period affect the *matching probabilities*, which serves as the *long-memory state* variable for the system. We provide conditions under which our setup suffices to yield perfect sorting or polarization of players to the two different locations based on group membership alone, and this statistical discrimination type of sorting is sustained by rational belief updating.

In our model under private monitoring, agents observe investment decisions and update beliefs about their own group or the other group depending on who they are matched with. But they must also consider what their match observed previously for the location decision and the investment game when they are old. To do that, they consider two factors: (1) how frequently their match met a member of either group and (2) what outcome they possibly experienced in that match. Both factors are endogenously determined. The first is governed by the law of likes meeting likes – a simple yet powerful first result of our paper wherein the likelihood of a meeting between two same group members in youth is higher than a meeting between a Blue and Red member. The second involves the formation of beliefs about the other players' updated beliefs. Together, players form endogenous perceptions.

To illustrate, suppose that players in the location choice stage of the game all anticipate that there will be more Blue members in the East. Consider the investment stage game history where a player observes Invest from a match with a Blue player in youth, a "good" outcome. If the player matched to this Blue player is also a Blue member, then, given the favorable history toward his own group, the Blue player finds it optimal to choose East to have a higher probability of being matched with another Blue member when old. On the other hand, if the player matched to the Blue member is a Red member, then, given the more favorable history of interaction with the Blue group member and the relatively unfavorable history toward his own group, the Red player finds it optimal to choose East to have a higher probability of being matched with a Blue group member when old. The former has the effect of widening the polarization of the two groups since there are already more Blue group members in the East, whereas the latter has the effect of reducing this gap. Yet, the former effect dominates the latter effect due to the natural law of likes meeting likes. Consider next the history where a player observes No Invest from a match with a Red player when young, a "bad" outcome. In fact, this history works in the same way as the history above: If the player matched to the Red player is a Blue player, then No Invest by a Red member is a relatively favorable history toward the Blue player's own group, but it is an unfavorable history toward his own group if the player matched to the Red member is

also a Red member. Both players in this case will choose East for the same reason, as well. In this case, however, the polarization-reducing effect dominates the polarization-widening effect due to the same law of likes meeting likes since a match between two Red members in youth is more frequent than a Red-Blue match.

Now, in order to assess which effect dominates, each player has to estimate the overall magnitude of each move, that is, what proportion of Blue group members and Red group members move to East or West, for which players use their own past experience. Hence, if the endogenous perception formation results in a larger probability of the Invest equilibrium relative to the probability of the No Invest equilibrium, then the location equilibrium yields a dynamical system in which the overall polarization-widening effect dominates. Yet, the convergence outcome is more *subtle*: The dynamical system may not result in polarization, despite the overall dominance of the polarizing effect since complete polarization requires that each period's polarized state surpasses that of the preceding period, creating an escalating trend. Without any exogenous force, achieving a progressively larger difference in each period is not possible; specifically, the *period-specific fixed point* of the system characterizing the population difference in the two locations is moving, and the state variable representing the difference in location choice probabilities can be smaller than the moving fixed point. Thus, in the absence of any external influence, the limiting outcome is a completely mixed state where both Red and Blue group members can be found in both locations.

However, suppose that only some proportion of agents makes endogenous location decisions each period; the remaining proportion follows exogenous, systematic polarization forces in their parent's locations where they were born. In that case, for a given amount of systematic polarization, there exists a corresponding initial population difference between the two locations such that in every period, the state variable representing the difference in location choice probabilities is always greater than the period's moving fixed point characterizing the population difference in the two locations.[4] The resulting dynamical system leads, in the limit, to a perfectly polarized society with all members of the Blue group located in one location and all members of the Red group located in the other location as opposed to the completely mixed state. The significance of this result is that even if some proportion of players is still making endogenous location decisions, systematic polarization overshadows their rational behavior, leading them to reach such an extreme completely polarized outcome; otherwise, the society converges to the completely mixed state.

Our paper is related to several different literatures. First, our paper is perhaps most closely related to an emerging literature on dynamic, forward-looking spatial models of migratory behaviour, *e.g.*, Allen and Donaldson (2020), and Kleinman, Liu and Redding (2023) or of voting behaviour, *e.g.*, Callander and Carbajal (2022). In these papers, investment, location or political positions depend on agents' expectations about the future spatial distribution of economic activity or of voter preferences. Callander and Carbajal (2022) in

[4]We restrict attention to a *robust* location strategy and provide conceptual problems without the robustness in order to rationalize it.

particular is illustrative and specifically addresses our question of whether and how a society may become more polarized over time. They consider a setting in which voters' preferences change in response to elections; specifically, voters update their ideal point closer to the ideal point of the party they voted for, but otherwise they vote rationally for the party closest to their ideal point. This preference updating, when combined with strategic best response behaviour by party candidates, creates a feedback loop that results in incrementally polarized voting outcomes. Under their dynamic, electoral competition moves away from efforts to capture swing voters and toward efforts to maximize the turnout of each party's supporters while avoiding losses of votes to abstention.

Our model is also characterized by a dynamical, feedback loop system that gives rise to polarization, but it is not preference-based, does not involve elections and is decentralized and not led by elites (party candidates); as such, our analysis is not restricted to polarization in the political domain or in migratory segregation, but can instead address polarization in socioeconomic status, geographical preferences, racial/ethnic identities, technological or educational divisions, *etc.*

Second, our paper is related to other approaches to understanding polarization. As noted earlier, this literature mainly considers *preference-based* explanations for sorting following the seminal work of Schelling (1971) and Tiebout (1956). There is also a literature on network-based explanations for sorting and polarization (see the survey by Jackson (2014) and the references therein) where agents can have special abilities to communicate or coordinate with agents who are members of their network. By contrast, our model has neither preferences for homophily nor any special communication channels that are exclusive to either group. In our environment, agents are "born innocent" without any biases for interacting with members of their own group or the other group, and it is mainly uncertainty about each group's type distribution and private monitoring, that results in the sorting of players by color – the observable identity – into the two different locations.

Regarding systematic polarization, it is well documented that social media can play a role in promoting and sustaining such polarization (see Zhuravskaya, Petrova and Enikolopov (2020) for a recent survey). We also capture the notion that social media can have an exogenous, amplifying effect for increasing interactions with members of one's own group, but in our setting, this amplification effect is systematic and applies equally to both groups. Overall, we view our results as providing *weaker* conditions for segregation or polarization than are obtained under assumptions of homophilic preferences or special group-specific communication or coordination facilities.

In this sense, our paper is also related to a research agenda in Sociology and Social Psychology that has sought to find the *minimal* conditions for the rise and maintenance of group identities. In one famous example, the "Robbers Cave" experiment of Sherif et al. (1961), 12-year boys were arbitrarily divided up into two groups at a summer camp and developed intense group identities and rivalries despite the fact that all of the boys were initially unknown strangers to one another and all came from similar middle-class backgrounds. The work of Sherif and associates led Tajfel (1974) and associates to develop the "minimal

group paradigm" of social psychology, an experimental protocol that seeks to understand in-group/out-group biases starting from the most minimal initial group conditions. The aim is to explicitly rule out preference-based explanations for intergroup discrimination, *e.g.*, due to prejudice, conflict, or stereotyping – as we do here as well – and thereby to understand the effects of minimal group assignment.

As Chen and Li (2009) have shown in experiments involving economic games, the use of this minimal group paradigm often suffices to generate large differences between the treatment of in- and out-group members.[5] In this paper, we also provide a model of how such in- and out-group distinctions can come about following in the *spirit* of the minimal group paradigm by making only an arbitrary initial group assignment to the players in our game, who are otherwise ex-ante identical, and we further show that polarization is not inevitable; it is also possible to have a completely mixed state as well, and we provide conditions under which either outcome can arise. That is, the focus of our paper is to identify conditions leading to polarized outcomes in an effort to better understand the question: What is the origin of polarization?

Theoretically, our paper is related to the matching literature, where the seminal work of Becker (1973) examines conditions under which an assortative matching arises as the equilibrium. However, the focus of this literature is on the stability of such equilibria (formally the core property); there is no strategic interaction between players in these models unlike in our approach. The matching model from this literature that is most closely related to our paper, Damiano and Li (2007), considers a *centralized* matching setup with two-sided incomplete information. In their model, a platform assigns agents to two different places (where they are randomly matched with one another within each place) to induce truth-telling and thereby achieves a second-best solution. By contrast, we consider a *decentralized* matching process but with strategic interaction between agents under *incomplete information*, which, more importantly, allows us to study a *dynamical procedure* on matching unlike the focus on the core. We also make use of the approach of Matsui and Matsuyama (1995) and Hofbauer and Sorger (1999) of using matching frictions and rational agents with perfect foresight in an explicitly dynamic setting to resolve equilibrium selection issues in bimatrix games.

Finally, our paper makes use of and advances the monotone comparative statics analysis of Milgrom and Shannon (1994). To show polarization, we need to make comparisons between homogeneous matches among members of the same group and heterogeneous matches among members of different groups. This is necessary for identifying a set of histories favorable toward a player's own group and to thereby generate a dynamical system. However, comparisons between different types of mappings are difficult using the standard monotone comparative statics approach. Therefore, we construct an auxiliary mapping with a param-

[5]There is also some non-experimental evidence that group sorting and identity is not entirely preference-based. Specifically, Kossinets and Watts (2009) examined 7,156,162 messages exchanged by 30,396 e-mail users at a large university over a 270-day period and found that similar individuals, *e.g.*, in terms of age, gender, field of study, location *etc.*, are more likely to communicate with one another than with others who are more different or distant.

eter in order to connect the two mappings in the spirit of Homotopy. The parameterized monotone comparative statics analysis is another separate contribution of this paper.

In the next section, we introduce our model. In Section 3, we present a simple two-period analysis. In Sections 4 and 5, we derive equilibrium conditions for the investment and endogenous location choice stages of our game. In Section 6, we characterize the population dynamics based on the results from the investment and location equilibrium and establish the main polarization results. Section 7 provides discussions and Section 8 concludes. All the proofs are collected in an appendix.

2 Dynamic spatial model

In this section, we develop a dynamic spatial model under group distributional uncertainty. There are two groups, Blue and Red, denoted by $g \in \{B, R\}$. Members of both groups live through two periods, youth and old age, and make three lifetime decisions. In each period, players of the same age, young or old, are randomly paired with one another at location ℓ which belongs to a finite set of locations L. Then, both young and old players participate in an investment game with their period match. In the interim period, players born at time $t-1$ make location decisions regarding their residence in old age.

Formally, given an initial population composition at $t = 0$, for $t = 1, 2, ...$, each individual born at period $t-1$ faces a decision problem in each of the following three phases:[6]

> Period $t-1$: Young members are born in their parent's location at time $t-1$ as members of the same group as their parent. These young players are then randomly paired with other young players of either group to play a stage game in the location of their birth.
>
> Interim period: A fraction of players born at time $t-1$ choose whether to move to a location ℓ, where they will reside in old age, period t.
>
> Period t: Old players of the $t-1$ generation are randomly paired with other old players of either group to play the stage game one final time in the location they chose for old age.

Note that while the young and the old play the stage game with members of their own generation, their matches within the same generation can still be *homogeneous or heterogeneous* with respect to group identities. That is, in each period t and every location ℓ, both homogeneous (same group) matches, (B, B), (R, R), and heterogeneous (or asymmetric group) matches, (B, R), are possible. The former is denoted by $m = S$, and the latter by $m = A$.[7]

In each period, each player chooses an action a to invest I or not invest N for a given match $m \in \{S, A\}$. This choice results in player i receiving a payoff $u(a_i, a_j, \theta_i)$, where θ_i (θ_j)

[6]The young player's parent does not necessarily imply a biological parent in a physical location. It refers to any older generation player of the same group identity who exerts significant influence (similar to

	Invest	No Invest
Invest	$d(\theta_i), d(\theta_j)$	$d(\theta_i) - 1, 0$
No Invest	$0, d(\theta_j) - 1$	$0, 0$

Table 1: The stage game

represents the ability or *type* of the row (column) player i (j). We can normalize this 2×2 game, without loss of generality, as shown in Table 1, assuming that $u(a_i, a_j, \theta_i)$ satisfies the *increasing difference* property. The function $d(\theta_i)$ can be interpreted as the normalized loss incurred by player i when that player unilaterally deviates from the action profile (I, I).[8] We assume that $d(\cdot)$ is continuously strictly increasing in order to induce a threshold equilibrium for the stage game for all $m \in \{S, A\}$.

A player's type, θ_i, is private information, while their group is perfectly identifiable. As in a typical Bayesian game, it is common knowledge that θ_i is drawn from a cumulative distribution (absolutely continuous) F_g for $g \in \{B, R\}$ with support $\Theta \equiv [\underline{\theta}, \overline{\theta}]$ given $\overline{\theta} > \underline{\theta}$. However, unlike the typical setup, no player in this model knows the true distribution. Consequently, there is uncertainty both about each group's distribution and about the type of a matched player j. Each group's distribution can be either F_X or F_Y, resulting in four possible combinations, $\{F_X, F_Y\} \times \{F_X, F_Y\}$, for the true distributions of the two groups (F_B, F_R). This modelling choice is essential for our perception-based polarization approach, as will be explained in detail below.

We aim to demonstrate the emergence of polarization from perception alone, without any initial biases. There are three conditions for unbiasedness. First, both groups must be of the same size, with unit mass. Second, while the increasing difference in the stage game implies strategic complementarity between players, unlike in preference-based polarization studies, the complementarity in this model exhibits *color-blindness*. This means that the degree of complementarity remains the same, irrespective of whether a player is matched with a member of the same group or the other group. Third, young players hold an *unbiased belief*: They believe that F_X and F_Y are equally likely for both groups, without using information on their own type (θ_i) to update their beliefs about their own group's distribution. This implies the belief of young players that each group's distribution is F_X is $\frac{1}{2}$. Specifically, the

a parental figure) over the young player, even in an online context.

[7]Our focus on intra-generational meetings – between young and between old members – enables us to utilize a symmetric equilibrium, even for matches between Blue and Red group member *i.e.*, heterogeneous matches, which makes our analysis tractable.

[8]A 2×2 game can be parameterized with $d_I(\theta_i) \equiv u(I, I, \theta_i) - u(N, I, \theta_i)$ and $d_N(\theta_i) \equiv u(N, N, \theta_i) - u(I, N, \theta_i)$, where $d_I(\theta_i)$ (resp. $d_N(\theta_i)$) represents the loss incurred by player i when that player unilaterally deviates from the action profile (I, I) (resp. (N, N)). This parameterization process is well known; *e.g.*, see Carlsson and van Damme (1993). Then, for the present model, we further normalize a 2×2 game as the stage game in Table 1 by defining $d(\theta_i) \equiv d_I(\theta_i)/(d_I(\theta_i) + d_N(\theta_i))$ if $d_I(\theta_i) + d_N(\theta_i) > 0$, where $d_I(\theta_i) + d_N(\theta_i) > 0$ iff $u(a_i, a_j, \theta_i)$ satisfies the increasing difference: $u(I, I, \theta_i) - u(I, N, \theta_i) > u(N, I, \theta_i) - u(N, N, \theta_i)$.

second and third unbiasedness – innocent young and "rational" old players – conditions will be clarified below in discussing the young player's investment stage equilibrium.

Each period's stage game is called the *investment stage*, and the interim phase's game is referred to as the *location stage*. Below, we explain each in detail; in particular, the perception formation process in Subsection 2.2.

2.1 Investment stage of the young

The equilibrium for young players, as well as that for the old, is a threshold for investing. Let k_j represent the threshold for player j such that player j chooses to invest if $\theta_j > k_j$ and chooses to not invest otherwise. For a given match $m \in \{S, A\}$ for the stage game in Table 1, taking into account the probability $\Pr(N|k_j)$ that player j chooses not to invest, the expected payoff to player i from investing (Invest) is given by:

$$U(\theta, k_j) \equiv d(\theta_i) - \Pr(N|k_j). \tag{1}$$

If player i chooses not to invest (NoInvest), his payoff is 0. Given this setup, a formal account of unbiasedness can now be provided. First, if we take a preference-based approach, a player's payoff $d^S(\theta_i)$ from a homogeneous match would differ from that $d^A(\theta_i)$ from a heterogeneous match. That is, unlike $d(\theta_i)$ in the young's payoff (1) and the stage game, preference-based approaches require $d^S(\theta_i) \neq d^A(\theta_i)$ for a positive measure. However, in our color-blind approach, the two payoffs are identical and equal to $d(\theta_i)$.

Further, if young players have biased beliefs in contrast to $\Pr(N|k_j)$ in (1), then we would have $\Pr^S(N|k_j) \neq \Pr^A(N|k_j)$ for some k_j. This situation is effectively *payoff equivalent* to the preference-based approach, contradicting our color-blind approach. In other words, for any color-blind payoff structure combined with biased beliefs, there exists a corresponding preference-based payoff structure with unbiased beliefs that yields exactly the same payoff for the young players. This payoff equivalence arises directly from the linearity of the young players' payoff in equation (1). As we adopt a color-blind payoff approach, it must be the case that $\Pr^S(N|k_j) = \Pr^A(N|k_j)$ as in the young's payoff in (1), which justifies our assumption of the young's unbiased beliefs.[9] With the unbiased belief that each group's type distribution is equally likely, in each location $\ell \in L$ and every matching $m \in \{S, A\}$, we have $\Pr(N|k_j) = \frac{1}{2}F_X(k_j) + \frac{1}{2}F_Y(k_j)$ for (1).

[9]That is, let us suppose $\Pr^S(N|k_j) \neq \Pr^A(N|k_j)$ and $\Pr(N|k_j)$ denotes the unbiased belief. Then, the color-blind payoff structure with the biased belief yields the expected payoff $d(\theta_i) - \Pr^m(N|k_j)$. Now, consider a preference-based approach by defining $d^m(\theta_i) \equiv d(\theta_i) - \Pr^m(N|k_j) + \Pr(N|k_j)$. It is easy to check that this preference-based approach with the unbiased belief yields the same payoff as that of the color-blind payoff structure with the biased belief above. Further, technically, if young players use own types, the posterior is given as $\Pr(X|\theta_i) = \frac{\Pr(\theta_i|X)\frac{1}{2}}{\Pr(\theta_i|X)\frac{1}{2}+\Pr(\theta_i|Y)\frac{1}{2}}$, where $\Pr(\theta_i|r) = f_r(\theta_i)$ PDF of $F_r(\theta_i)$ for $r \in \{X, Y\}$. Then, $\Pr(X|\theta_i) - \frac{1}{2} = \frac{f_X(\theta_i) - f_Y(\theta_i)}{2[f_X(\theta_i)+f_Y(\theta_i)]}$. Now, note that only with F_X FOSD F_Y – as we have in the present model – both $f_X(\theta_i) > f_Y(\theta_i)$ and $f_X(\theta_i) < f_Y(\theta_i)$ can arise, as well known. Hence, even for a higher θ_i, if $f_X(\theta_i) < f_Y(\theta_i)$, player i thinks that the group's distribution is *less* optimistic, so he is less likely to choose Invest. Thus, we cannot guarantee monotonicity to enable us to have a threshold equilibrium.

For the investment stage equilibrium, we rely on the following assumptions.

(A1) For each $r \in \{X, Y\}$, if F_r is the true distribution, there exists an interior equilibrium threshold, denoted by k_r.
(A2) There exists a subinterval $\Gamma \subseteq \Theta$ such that $\Gamma \equiv \{\theta \in \Theta : F_X(\theta) < F_Y(\theta)\}$.
(A3) For each pair $\theta' > \theta$ in Γ, $d(\theta') - d(\theta) \geq \frac{1}{2}[F_X(\theta') - F_X(\theta)] + \frac{1}{2}[F_Y(\theta') - F_Y(\theta)]$.

By restricting the class of distributions that satisfy (A1), we guarantee *interior* solutions for the young and old player's equilibrium in subsequent sections, and for (A2), we use the first-order stochastic dominance (FOSD) in the local sense, including the standard FOSD, to define F_X as the "better" distribution on a subinterval of the support.[10] We make assumption (A3) for two reasons that concern the young's equilibrium outcomes, which turns out to be useful for the old player's equilibrium outcomes as well. First, by (A3), the stochastically dominant distribution, F_X, yields a *lower* threshold for Invest, $k_X < k_Y$, meaning that there is a higher probability of choosing Invest for $r = X$ if it were known as the true distribution.[11] We may call $[k_X, k_Y]$ the *effective support* in the sense that all equilibrium thresholds arise in that interval.[12] Second, by (A1), there exists an interior equilibrium threshold k for the young such that

$$d(k) = \frac{1}{2}F_X(k) + \frac{1}{2}F_Y(k). \tag{2}$$

Then, by (A3), there exists a *unique* k, and further, with the monotone mapping $d^{-1}\big(\frac{1}{2}F_X(k) + \frac{1}{2}F_Y(k)\big)$, for a change from (X, Y) to (X', Y') such that $F_{X'}(\theta) + F_{Y'}(\theta) < F_X(\theta) + F_Y(\theta)$, a comparative statics analysis yields that k *decreases*. Each of these two outcomes above is intuitive.

For concreteness, Figure 1 shows two example distributions, $F_X(\theta) = \theta$ and $F_Y(\theta) = \frac{1-e^{-\lambda\theta}}{1-e^{-\lambda}}$ for $\theta \in [0, 1]$, together with d that satisfy assumptions (A1)-(A3).

---

[10]For each $r \in \{X, Y\}$, if F_r were known to be the true distribution, then Invest yields player i the expected payoff $d(\theta_i) - F_r(k_j)$, so $F_r(k_j)$ is the expected probability that player j does not invest. By (A1), there exists an interior equilibrium threshold, k_r, satisfying $d(k_r) = F_r(k_r)$.

[11]Note that (A3) implies that for any pair $\theta' > \theta$ in Γ, $d(\theta') - d(\theta) \geq F_r(\theta') - F_r(\theta)$ for at least one $r \in \{X, Y\}$. Now, suppose, on the other hand, $k_X \geq k_Y$. Then, by (A1)-(A2), $d(k_X) - d(k_Y) = F_X(k_X) - F_Y(k_Y) < F_r(k_X) - F_r(k_Y)$ for all r, which yields a contradiction for both $k_X = k_Y$ and $k_X > k_Y$; in particular, the latter contradicting (A3).

[12]See Yoo (2014) for this and other more technical aspects regarding the existence of such an equilibrium; for instance, a condition for the support of the distributions can be sufficient for assumption (A1). The analysis in that paper applies to a setting with a single group distribution and matching within that single group.

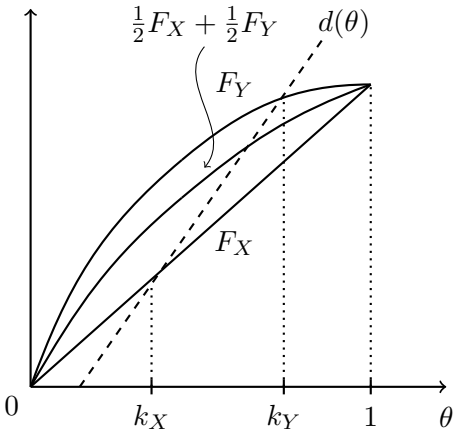

Figure 1: $d(\theta)$ and F_r satisfying (A1)-(A3)

2.2 Location stage

After the game, each young player can observe the action chosen by their matched player j.[13] The history of play among the young not only influences each player's investment decision in old age but also impacts each player's location decision in the interim period, considering the expected future payoff from matching with an old player.

Each player's belief updating follows the standard Bayes' rule. Let $\pi(I)$ and $\pi(N)$ represent an old player's belief regarding the probability that the distribution of a matched partner's group is F_X given that the old player previously observed an investment action (I) or a non-investment action (N), respectively. By Bayes' rule, an old player's beliefs are updated as follows:

$$\pi(I) = \frac{(1 - F_X(k))}{(1 - F_X(k)) + (1 - F_Y(k))} \text{ and } \pi(N) = \frac{F_X(k)}{F_X(k) + F_Y(k)}. \tag{3}$$

Then, a player who had a good experience with a member of group $g \in \{B, R\}$ can become more optimistic about that group's distribution. If there was no prior match with a member of the group, which is denoted by $\emptyset$, then there is no update for that group. Players use these updated beliefs to make location choices in the interim period as well. We will use a group-specific superscript I^g, N^g, or $\emptyset^g$ for $g \in \{B, R\}$ if doing so helps us keep track of a history with respect to a certain group.

For study of the role of perception, we let each player observe only his own history – *i.e.*, private monitoring not public monitoring. This means that first, after observing a history, players update their beliefs about the distribution of group g via Bayes' rule in (3), and next, they also have to form beliefs about the other players' updated beliefs. That is, the belief

[13]They, however, cannot directly observe the type θ_j of their partner. If a young player can observe the type θ_j of their partner and update their beliefs about the other group, we cannot guarantee monotonicity in order to establish a threshold equilibrium, for the same reason as explained in Footnote 9.

formation of players in this model involves higher order. However, there is no signal structure exogenously given in this model; all observations are endogenously generated. This marks a departure from other higher order belief models. Hence, in this *endogenous* perception formation, each player forms beliefs not just about what signal the other players observed but also about how those signals were generated endogenously. This key aspect of the model affects the investment stage equilibrium of the old players as well as the location stage equilibrium

In addition, we introduce a friction in players' location choices in the spirit of Matsui and Matsuyama (1995). In their paper, at each point in continuous time, every player must make a commitment to a pure strategy, but the opportunities to switch between actions arrive randomly, following a Poisson distribution. By contrast with Matsui and Matsuyama (1995), the present model adopts a discrete time framework in order to focus on learning under distributional uncertainty. Hence, a discrete counterpart of the continuous time approach is that in each time $t = 1, 2, ...,$ only $\alpha \in (0, 1)$ proportion of players in each group are randomly chosen to make location choices. Hence, for the location stage equilibrium, we add the following two assumptions.

(A4) Monitoring is private.
(A5) In each time t, a fraction $\alpha \in (0, 1)$ of each group $g \in \{B, R\}$ makes rational location choices.

The private monitoring assumption (A4) means that each player must anticipate his partner's history to obtain a "perception about the other's endogenous perception" about a group. With the addition of this friction (A5), as in other search models, we can select a reasonable equilibrium in this dynamic spatial model, despite the conceivable multiplicity-of-equilibrium problem in a belief-driven model. The endogenous perception formation requires two elements; how frequently a group g player is matched with a group g' player and what histories they can observe in the stage game.

To answer the first question, we find the overall probability q_t that a player is matched with a member of his own group. Without loss of generality, we consider two locations, East (E) and West (W) in what follows.[14] We let $\mathcal{P}_t^B$ ($\mathcal{P}_t^R$) denote the probability that B (R) group members move to location E. Then, the proportion of B members in E is $\frac{\mathcal{P}_t^B}{\mathcal{P}_t^B+\mathcal{P}_t^R}$ and the proportion of B members in the other location W is $\frac{1-\mathcal{P}_t^B}{2-\mathcal{P}_t^B-\mathcal{P}_t^R}$. Together, the overall probability q_t that a player is matched with a member of their own group in both locations is given by

$$q_t = \frac{\mathcal{P}_t^B + \mathcal{P}_t^R - 2\mathcal{P}_t^B\mathcal{P}_t^R}{(\mathcal{P}_t^B + \mathcal{P}_t^R)(2 - \mathcal{P}_t^B - \mathcal{P}_t^R)}. \tag{4}$$

This formula yields a simple but powerful first result, the lemma below that we term "Natural law of likes meeting likes." According to this lemma, homogeneous matchings with a member

[14] In Appendix B.1, we show that our results can be generalized to $L \geq 2$ finite locations with members of the two groups congregating in two different sets of locations.

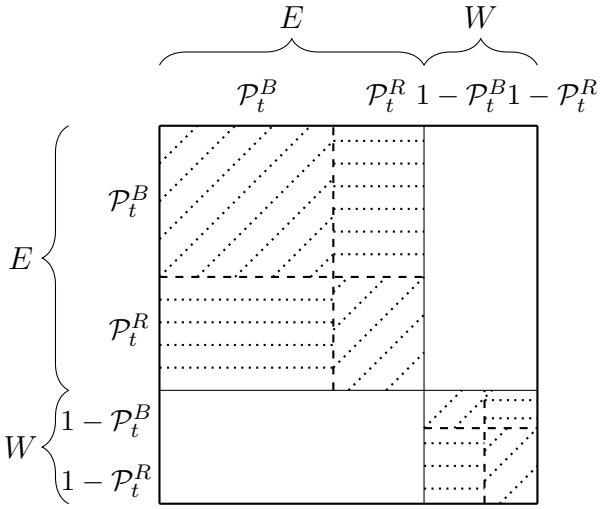

Figure 2: Same & different group matching probabilities

from the same group are more likely than heterogenous matchings with a different group member, so long as $\mathcal{P}_t^B$ and $\mathcal{P}_t^R$ are different.

Lemma 1 *(Natural law of likes meeting likes) The homogeneous group matching probability q_t is greater than the heterogeneous group matching probability $1-q_t$, if $\mathcal{P}_t^B \neq \mathcal{P}_t^R$.*

The law can be straightforwardly understood. By simple algebra, we have $q_t^B + q_t^R \geq (>)\, 2q_t^{BR} = 2q_t \geq (>)\, 2q_t^{BR}$ for all $\mathcal{P}_t^B$, $\mathcal{P}_t^R$ ($\mathcal{P}_t^B \neq \mathcal{P}_t^R$).[15] This is illustrated in Figure 2, where the shaded area with diagonal lines depicts $q_t^B + q_t^R$, whereas the area with horizontal lines depicts $2q_t^{BR}$, and the sum of the former is greater than the sum of the latter except in the case where $\mathcal{P}_t^B = \mathcal{P}_t^R$. In other words, there is an underlying force that *generically* makes matches with members of the same group more frequent.

A second important element for determining the proportion of players observing each history is the following probabilities:

$$\begin{aligned} p_I(\omega^g) &\equiv \pi(\omega^g)(1-F_X(k)) + (1-\pi(\omega^g))(1-F_Y(k)), \\ p_N(\omega^g) &\equiv \pi(\omega^g)F_X(k) + (1-\pi(\omega^g))F_Y(k). \end{aligned} \tag{5}$$

Hence, $p_I(\omega^g)$ (resp. $p_N(\omega^g) = 1 - p_I(\omega^g)$) denotes a player's beliefs about the probability that Invest (resp. NoInvest) was generated by a group g when young, given the player's own history ω^g about group g. The assumption of unbiased beliefs requires that young players adopt the same equilibrium threshold k in (2). Then, conditional on $r \in \{X, Y\}$, the

[15] The same group matching probability from a B member's point of view is $q_t^B = \frac{(\mathcal{P}_t^B)^2}{\mathcal{P}_t^B+\mathcal{P}_t^R} + \frac{(1-\mathcal{P}_t^B)^2}{2-\mathcal{P}_t^B-\mathcal{P}_t^R}$ and that from an R member's point of view is $q_t^R = \frac{(\mathcal{P}_t^R)^2}{\mathcal{P}_t^B+\mathcal{P}_t^R} + \frac{(1-\mathcal{P}_t^R)^2}{2-\mathcal{P}_t^B-\mathcal{P}_t^R}$, both of which yield (4), that is, $q_t^B = q_t^R = q_t$. On the other hand, the different group matching probability from either group member's point of view is $q_t^{BR} = \frac{\mathcal{P}_t^B\mathcal{P}_t^R}{\mathcal{P}_t^B+\mathcal{P}_t^R} + \frac{(1-\mathcal{P}_t^B)(1-\mathcal{P}_t^R)}{2-\mathcal{P}_t^B-\mathcal{P}_t^R}$.

probability that Invest or NoInvest was observed in the young player's stage game is identical for both types of match so that $\Pr(\text{Invest}|r) = 1 - F_r(k)$ for $r \in \{X, Y\}$ or $\Pr(\text{NoInvest}|r) = F_r(k)$ for $r \in \{X, Y\}$. While the unbiased-beliefs assumption makes the analysis simple, the real complications arise as to how one should evaluate the likelihood of each group's distribution when they are in the interim stage and old – perceptions about endogenous perceptions. If a player had a positive experience with a member of group g when young, then this would make him more optimistic about that group, *i.e.*, $\pi(I) > \pi(N)$ by (3). It would also increase (resp. decrease) his beliefs that others had a good (resp. bad) experience – a higher (resp. lower) probability of observing Invest (resp. NoInvest), *i.e.*, $p_I(I) > p_I(N)$ (resp. $p_N(I) < p_N(N)$) – with this same group as well, despite the same k threshold for Invest in either homogeneous or heterogeneous matches in youth. Together, the two components (4) and (5) constitute the endogenous perception formation process of this model.

While anticipating a future partner's history, each player also plans a location choice. To simplify notations, we consider the location choice problem from a group B player's perspective. As such, a history is written as $\omega = (\omega^B, \omega^R)$ with a fixed order, which yields a set of histories $\Omega \equiv \{(I, \emptyset), (\emptyset, N), (N, \emptyset), (\emptyset, I)\}$ where, for example, $(I, \emptyset)$ means that a player observes Invest from a B player when they are young. In addition, with the location stage friction, each player's location decision can be constrained by a previous period's population difference $\Delta\mathcal{P}_{t-1} \equiv \mathcal{P}^B_{t-1} - \mathcal{P}^R_{t-1}$. We simply denote the difference by $z_{t-1} \in [-1, 1]$ if necessary. By allowing for some proportion of *exogenous* location choices, based on the idea of Matsui and Matsuyama (1995), we can construct a Markov location strategy where each player's period t location strategy depends not only on their beliefs about which location has more of their own group members but also on the period $t-1$ actual population difference, which yields a more reasonable and robust equilibrium prediction.

Then, a location strategy of group g is given by

$$\ell^g_t : \Theta \times \Omega \times [-1, 1] \to L, \tag{6}$$

where note that we can focus on a pure strategy location equilibrium without loss of generality.[16] In Section 5, we show that the *optimal* ℓ^g_t in fact only depends on ω and $\Delta\mathcal{P}_{t-1}$, not a player's intrinsic type θ. Given (ω, ℓ^g_t), let $P^B_t(\omega, \ell^B_t)$ denote a player's beliefs with history ω about the proportion of B players located in E among B members in period t, and $P^R_t(\omega, \ell^R_t)$ denote his beliefs with history ω about the proportion of R players located in E among R members in period t. With the friction, they can be derived as

$$\begin{aligned} P^B_t(\omega, \ell^B_t) &= \alpha\mathbb{E}\big[\mathbf{1}_{\{\ell^B_t(\widetilde{\theta}, \widetilde{\omega}, z_{t-1}) = E\}} \mid \omega\big] + (1-\alpha)\mathcal{P}^B_{t-1}, \\ P^R_t(\omega, \ell^R_t) &= \alpha\mathbb{E}\big[\mathbf{1}_{\{\ell^R_t(\widetilde{\theta}, \widetilde{\omega}, z_{t-1}) = E\}} \mid \omega\big] + (1-\alpha)\mathcal{P}^R_{t-1}, \end{aligned} \tag{7}$$

[16] If there exists a mixed location strategy, it does *only* when $\Delta P_t(\omega, \ell_t) = 0$ in (8); it does not matter where to move to (see Section 5). In Appendix B.2, we show that there is no robust mixed strategy location equilibrium. Also, for the investment stage game of the young and the old, we consider a threshold equilibrium with no loss of generality, as well known.

where a proportion α of each group makes the endogenous location decisions, whereas the remaining $1-\alpha$ proportion does not change their locations (we use z_{t-1} to denote $\Delta\mathcal{P}_{t-1}$).[17]

Note that given (4) and (5), $P_t^B(\omega, \ell_t^B)$ and $P_t^R(\omega, \ell_t^R)$ in (7) can be derived. That is, using them, each player first forms his perception about group g, and then perception about the other's perception. To illustrate, let us fix a player's history as ω. Then, based on this history $\omega = (\omega^B, \omega^R)$, the player updates $\pi(\omega^g)$, and then assesses each probability that an arbitrary B player or R player observes a history $\widetilde{\omega}$ that belongs to the set $\Omega \equiv \{(I, \emptyset), (\emptyset, N), (N, \emptyset), (\emptyset, I)\}$. To make this concrete, for instance, consider $\widetilde{\omega} = (N, \emptyset)$ that an arbitrary B player could have among the set Ω. Then, the player reasons that with probability q_{t-1} in (4), an arbitrary B player met a member of his own group in youth, and with probability $p_N(\omega^B)$ in (5), an arbitrary B player observed N from his own group member. Now, if a B player with $\widetilde{\omega} = (N, \emptyset)$ moves to E, then that population proportion $q_{t-1}p_N(\omega^B)$ must be counted for in $\mathbb{E}\big[\mathbf{1}_{\{\ell_t^B(\widetilde{\theta}, \widetilde{\omega}, z_{t-1}) = E\}} \mid \omega\big]$ for $P_t^B(\omega, \ell_t^B)$ (for $P_t^R(\omega, \ell_t^R)$, one can use a similar procedure). Each of these two expectations in (7) is taken over four possible histories $\widetilde{\omega} \in \Omega$ that an arbitrary group g old player could have, conditional on a player's perspective with his own history ω. We emphasize that for the player's reasoning, he uses his own *private experience* $\omega = (\omega^B, \omega^R)$ to calculate the probability of an arbitrary B observing N; that is, $q_{t-1}p_N(\omega^B) = q_{t-1}p_N(I)$ if $\omega^B = I$ and $q_{t-1}p_N(\omega^B) = q_{t-1}p_N(N)$ if $\omega^B = N$. One can obtain intuition of this endogenous perception formation process from Section 3 and its formality from the proof of Proposition 1.

The belief difference between $P_t^B(\omega, \ell_t^B)$ and $P_t^R(\omega, \ell_t^R)$ plays a critical role in both group members' location decisions, which is denoted by

$$\Delta P_t(\omega, \ell_t) \equiv P_t^B(\omega, \ell_t^B) - P_t^R(\omega, \ell_t^R), \tag{8}$$

where we denote $\ell_t \equiv (\ell_t^B, \ell_t^R)$. If a player believes that a location has more of his group members and has a positive experience with his own group member, then he will choose that location optimally. Thus, the location choice depends on the belief difference $\Delta P_t(\omega, \ell_t)$ and the endogenous perception formation.

With no friction, *i.e.*, $\alpha = 1$, the next period beliefs $\Delta P_t(\omega, \ell_t)$ does not depend on a previous real population difference $\Delta\mathcal{P}_{t-1}$ at all. Now, with some friction $\alpha \in (0, 1)$, we say that an optimal location strategy ℓ_t is *robust* if there exists a sufficiently small open interval I around α such that the strategy ℓ_t is optimal for all $\alpha' \in I$. For a given size of friction α combined with an external influence, we aim to find out an initial population composition that yields complete polarization as the limiting outcome of the state variable $\Delta\mathcal{P}_{t-1}$, so if an optimal strategy ℓ_t cannot survive for a sufficiently small perturbation of α, the result will not be robust. Equally importantly, one can find serious conceptual problems with no robustness applying to location strategies in the end of Section 3 and Section 5, above the related key finding, Proposition 5.

[17] As in the formula, $P_t^g(\omega, \ell_t^g)$ depends on $\mathcal{P}_{t-1}^B$ as well as q_{t-1}, but to avoid burdensome notations, we keep it simple as $P_t^g(\omega, \ell_t^g)$ despite their apparent roles in the subsequent analyses.

2.3 Investment stage of the old

The old players update their beliefs about the distributions of both groups following the Bayes' rule (3) based on their experience when they are young. Then, an old player's payoff has a payoff structure similar to (1), but the probability that his matched player j chooses NoInvest depends on his or her history of play from the investment stage game when young and the belief updating in (3).

This means that unlike the young players' strategy in (2), despite the three unbiasedness conditions including, particularly, the color-blind payoff structure, the old players' strategy depends on both a type of meeting and a history, that is, whether it is a homogeneous and heterogeneous match, so solving the problem requires a multidimensional fixed point. We denote a set of histories relevant for a type of meeting $m \in \{S, A\}$ by Ω^m with its element ω^m. Then, an old player's strategy at time t for a type of meeting m is a mapping such that

$$s_t^m : \Theta \times \Omega^m \to \{I, N\}, \tag{9}$$

where note that this strategy comes with a time t subscript, in contrast to the young player's threshold k, since the $t-1$ period's same group matching probability q_{t-1} in (4) can also affect the old player's strategy in anticipating his opponent's past experience.

In a homogeneous match, a player's beliefs about his group conditions beliefs of the matched old partner from the same group too. However, in a heterogeneous match, even conditional on a B group member's beliefs about the R group, beliefs of the matched old R partner about the B group still remains, which demands again perception about the other's endogenous perception. Finding the *history-dependent* homogeneous and heterogeneous equilibria by itself is the subject of intensive study in Section 4. All in all, what is critical for the intriguing point in terms of the old's stage game as well as the location equilibrium is not just how each player evaluates the youthful partner's group but also how each player thinks the future partner will evaluate his own group.

2.4 Equilibrium

We explain the definition of the dynamic spatial equilibrium in words first, saving the formal definition for later in Section 5. The full analysis, developed later in Sections 4 and 5, requires the following standard backward induction. First, the investment stage for old players yields two equilibrium payoffs – one for matching with a member of the same group and another for matching with a member of the other group. Then, in the location stage, anticipating the next-period population ratios in the different locations, players in the interim stage decide where to move to by comparing the future benefits of homogeneous versus heterogeneous matches when old. For both equilibria, each player relies on how he thinks his future partner will evaluate his own group – perceptions about perceptions – as well as how he evaluates his youthful partner's group. This belief updating is based on what each player observes in the young's equilibrium.

The next section aims to build intuition for our dynamical system, that can give rise to either polarized or mixed outcomes.

3 A simple two-period analysis

In this section, we present a simplified two-period analysis. The main purpose of this section is to demonstrate how perceptions about endogenous perceptions can be a driving force in both the investment stage equilibrium of the old players and the location stage equilibrium. For them, we need to determine (i) what histories are favorable toward a player's own group and (ii) who observes each history and moves to each location.

Analysis of a full-fledged model in this section would be too complicated to convey the intuition, so we focus on the second part (ii) in this section. The first part (i) is based on the payoff comparison between homogeneous matches and heterogeneous matches when old, and just to give a glimpse of (i), suppose that a B player observed Invest from another young B. Then, it would make him more optimistic about his own group in a homogeneous match when they are old. Yet, even when an old B player is matched with an old R member, the positive experience with a young B member would make the old B member think that the old R member was more likely to observe the same good behavior from a young B member in the past as well. This heterogeneous match yields a higher payoff to the B player with the same history too. Hence, the comparison comes down to which positive effect is stronger (see Sections 4 and 5 for the details). The case for an R group member can be symmetrically described. We later show in Section 5 that given $\Omega \equiv \{(I,\emptyset),(\emptyset,N),(N,\emptyset),(\emptyset,I)\}$, the characterization of the B and R player's histories that are either favorable (+) or unfavorable (-) toward members of their own group yields

$$\begin{aligned} \Omega^{B+} &= \{(I,\emptyset),(\emptyset,N)\} \text{ and } \Omega^{B-} = \{(N,\emptyset),(\emptyset,I)\}, \\ \Omega^{R+} &= \{(\emptyset,I),(N,\emptyset)\} \text{ and } \Omega^{R-} = \{(\emptyset,N),(I,\emptyset)\}. \end{aligned} \tag{10}$$

Then, if a player *believes* that a location has more of his own members, in the interim stage, he finds it optimal to go to that location when the future homogeneous payoff is greater than the future heterogeneous one. To illustrate the perception formation process for the second part (ii), let us call a player Abe and fix his history as $\omega \in \Omega$. With this ω, Abe contemplates the probability that others observe each $\widetilde{\omega} \in \Omega$. For instance, consider an arbitrary B player Betty and further $\widetilde{\omega} = (N,\emptyset)$ as her history among the four possible histories. Since Betty met a member of her own group in youth with probability q_{t-1}, and Betty observed N with probability $p_N(\omega^B)$, the probability that a B player observes history $\widetilde{\omega}$ from Abe's perspective given ω can be calculated as $q_{t-1}p_N(\omega^B)$. Now, in assessing $p_N(\omega^B)$, Abe uses ω^B from his own experience $\omega = (\omega^B, \omega^R)$. That is, if Abe's history is (I,ϕ), then (Abe's belief about) the probability that Betty observes $\widetilde{\omega} = (N,\emptyset)$ is $p_N(I)$; if Abe's history is (ϕ,I), then the probability that Betty observes $\widetilde{\omega} = (N,\emptyset)$ is $p_N(\phi)$; and so on. In other words, players with different histories have different assessments of the probability

	$\Delta P_t(\omega, \ell_t) > 0$ for all ω	$\Delta P_t(\omega, \ell_t) < 0$ for all ω
$\omega \in \Omega^{B+} = \Omega^{R-}$	B (R) chooses E	B (R) chooses W
$\omega \in \Omega^{B-} = \Omega^{R+}$	B (R) chooses W	B (R) chooses E

Table 2: The binary splitting location strategy

that Invest or NoInvest was generated for each group. As illustrated, the reasoning relies on $p_I(\omega^g)$ and $p_N(\omega^g)$ in (5) given $\omega = (\omega^B, \omega^R)$ as well as q_{t-1} in (4). Hence, $p_I(\omega^g)$ (resp. $p_N(\omega^g)$) is a player's belief with a history $\omega = (\omega^B, \omega^R)$ about the probability that a young group g player generated Invest (resp. NoInvest) in a match. These *beliefs* depend on how a particular personal experience ω in youth makes the player optimistic about group g in the interim stage and old age – *i.e.,* group g more likely to be from a good distribution via $\pi(\omega^g)$, and Invest was more likely to be generated by that group.

Now, we show that a location strategy is an equilibrium given the history sets for group B and R in (10), while leaving the uniqueness of the location equilibrium to Section 5. A location equilibrium consists of two components: optimality and consistency (see Section 5 for a formal definition). Consider a simple strategy in Table 2, called a binary splitting location strategy. One can find that it satisfies the optimality condition because those players with histories that belong to the favorable sets find it optimal to go to the location where there are more members of their own group. Regarding the consistency, considering the optimal strategies chosen by all players with different histories, their initial expectations about the proportion of group members in each location must be shown to be correct.

Finding $P_t^B(\omega, \ell_t^B)$ and $P_t^R(\omega, \ell_t^R)$ involves the aforementioned delicate procedure: Each player uses his own history to anticipate the histories that the others may have. To focus on this endogenous generation of beliefs, in this section, we consider the only endogenous location decision; that is, let $\alpha = 1$ in (7). Now, to show that the binary splitting location strategy in Table 2 is an equilibrium, we start by supposing that $\Delta P_t(\omega, \ell_t) > 0$ in (8); that is, there are more B members in the East in equilibrium. Consider a history $(I, \emptyset)$ that may have been experienced by others, which belongs to Ω^{B+} and Ω^{R-}; that is, the case where a player has a favorable history toward his own group if the player is a B member but an unfavorable history toward his own group if an R member. Anticipating that there will be more B members in the East, a B member with that history finds it optimal to choose E to meet with members of his own group when old, and an R member with that history finds it optimal to go to E as well in order to meet with members of the other group. Specifically, the former case arises when with probability q_{t-1}, a B player met with a member of his own group, and with probability $p_I(\omega^B)$, he observed Invest in youth, whereas the latter case arises when with probability $1 - q_{t-1}$, an R player met with a B player and observed Invest in youth with $p_I(\omega^B)$. Then, two different probabilities of observing I from a B member can

be derived as[18]

$$\begin{cases} q_{t-1}p_I(\omega^B) & \text{if a } B \text{ member met another } B \text{ in youth observing } (I,\emptyset), \\ (1-q_{t-1})p_I(\omega^B) & \text{if an } R \text{ member met a } B \text{ in youth observing } (I,\emptyset). \end{cases}$$

Again, the above probabilities about others' histories depend on who is assessing those histories – the one's history ω^B from $\omega = (\omega^B, \omega^R)$. The overall effect of observing I from a B member is positive due to the law of likes meeting likes by Lemma 1 such that $(2q_{t-1} - 1)p_I(\omega^B) > 0$ – the former match (B, B) arises with a higher probability in youth. The positive sign means that with this history, the proportion of B members moving to the East is greater.

Now, consider another history of the same kind $(\emptyset, N)$, which also belongs to Ω^{B+} and Ω^{R-}; for this history, players want to locate in the East as well since the history indicates a bad outcome from meetings in youth with members of the R group. Similarly, two different probabilities of observing N from an R member can be derived as

$$\begin{cases} (1-q_{t-1})p_N(\omega^R) & \text{if a } B \text{ member met an } R \text{ in youth observing } (\emptyset,N), \\ q_{t-1}p_N(\omega^R) & \text{if an } R \text{ member met another } R \text{ in youth observing } (\emptyset,N). \end{cases}$$

In this case, however, the overall effect of observing N from an R member is *negative*, $-(2q_{t-1}-1)p_N(\omega^R) < 0$, due to the same law of likes meeting likes – the latter match (R, R) arises with a higher probability in youth. The negative sign means that with this history, the proportion of R members moving to the East is greater.

Thus, the dynamics embody two contrasting forces: The first history increases polarization – since there are already more B members in the East – while the second diminishes it. This implies that to satisfy the consistency, *i.e.*, the belief $\Delta P_t(\omega, \ell_t) > 0$ is correct, we need a larger probability of observing Invest from a B member relative to the probability of observing No Invest from a R member, which requires a condition as below. The above two cases yield $P_t^B(\omega, \ell_t^B)$ and $P_t^R(\omega, \ell_t^R)$ for this section such that

$$\begin{aligned} P_t^B(\omega, \ell_t^B) &= q_{t-1}p_I(\omega^B) + (1-q_{t-1})p_N(\omega^R), \\ P_t^R(\omega, \ell_t^R) &= (1-q_{t-1})p_I(\omega^B) + q_{t-1}p_N(\omega^R). \end{aligned} \tag{11}$$

Then, by taking the difference between $P_t^B(\omega, \ell_t^B)$ and $P_t^R(\omega, \ell_t^R)$ above – or by summing the two differences above, $(2q_{t-1}-1)p_I(\omega^B)$ and $-(2q_{t-1}-1)p_N(\omega^R)$, up – we have the first

[18]The probability of observing any history when players are young is not location-specific, precisely because all the young members choose the same k. For instance, $q_{t-1}p_I(\omega^B)$ is given by

$$\mathcal{P}_{t-1}^B \left(\frac{\mathcal{P}_{t-1}^B}{\mathcal{P}_{t-1}^B + \mathcal{P}_{t-1}^R} \right) p_I(\omega^B) + (1 - \mathcal{P}_{t-1}^B) \left(\frac{1 - \mathcal{P}_{t-1}^B}{2 - \mathcal{P}_{t-1}^B - \mathcal{P}_{t-1}^R} \right) p_I(\omega^B) = q_{t-1}p_I(\omega^B),$$

where the first term is the same group matching probability times the probability of observing I in the East, whereas the second is that in the West.

branch of the formula for the belief dynamics below. The second branch, symmetrically, can be derived now based on the opposite beliefs $\Delta P_t(\omega, \ell_t) < 0$. Together, we have

$$\Delta P_t(\omega, \ell_t) = \begin{cases} (2q_{t-1} - 1)[1 - p_N(\omega^B) - p_N(\omega^R)] & \text{if } \Delta P_t(\omega, \ell_t) > 0, \\ -(2q_{t-1} - 1)[1 - p_N(\omega^B) - p_N(\omega^R)] & \text{if } \Delta P_t(\omega, \ell_t) < 0. \end{cases} \tag{12}$$

By the natural law of likes meeting likes (Lemma 1), the same group matching probability is higher; $2q_{t-1} - 1 > 0$. Hence, to induce the correct expectations requires a condition $p_I(\omega^B) - p_N(\omega^R) = 1 - p_N(\omega^B) - p_N(\omega^R) > 0$. As such, the *sum of a player's beliefs about the probability* of the others observing NoInvest plays a key role, which is defined as

$$\widehat{p}_N(\omega) \equiv p_N(\omega^B) + p_N(\omega^R), \tag{13}$$

where one can find that the vector's order does not matter, *i.e.*, $\widehat{p}_N(\omega) = \widehat{p}_N(\omega')$ for $\omega = (I, \emptyset)$, $\omega' = (\emptyset, I)$, and that $\widehat{p}_N(I, \emptyset) < \widehat{p}_N(N, \emptyset)$. Note that if there is no perception about perception (behave like a naive young player), then old players will have exactly the same payoff regardless of whether they meet the same group member or not. Hence, it does not matter where to move to, so the perception formation is an integral part of our analysis.

The perception formation process, however, is not sufficient to derive a formula for the dynamical system. Another key aspect for derivation of the formula is the law of likes meeting likes in (4). We rewrite it here as $q_{t-1} = \frac{1}{2} + \frac{\Delta \mathcal{P}^2_{t-1}}{2A(2-A)}$ or $(2q_{t-1} - 1) = \frac{\Delta \mathcal{P}^2_{t-1}}{A(2-A)}$ so that it can be incorporated into (12), where we denote by $A \equiv \mathcal{P}^B_{t-1} + \mathcal{P}^R_{t-1}$ the sum of the two moving probabilities. Then, by reformulating (12), we obtain *belief population composition dynamics* such that

$$\Delta P_t(\omega, \ell_t) = \begin{cases} \frac{1-\widehat{p}_N(\omega)}{A(2-A)} \Delta \mathcal{P}^2_{t-1} & \text{if } \Delta P_t(\omega, \ell_t) > 0, \\ -\frac{1-\widehat{p}_N(\omega)}{A(2-A)} \Delta \mathcal{P}^2_{t-1} & \text{if } \Delta P_t(\omega, \ell_t) < 0. \end{cases} \tag{14}$$

The sum A is a constant in this section without the friction, since the q_{t-2} terms in $\mathcal{P}^B_{t-1}$ and $\mathcal{P}^R_{t-1}$ cancel out one another.[19] Later in Section 6, with the friction α, the sum A will have a time subscript without no such cancellation, which will generate a moving fixed point of our system. Now, extending (14), we derive the real dynamical system from the point of view of the modeler who *knows* the true F_B and F_R, essentially by transforming beliefs about the probability of the others observing NoInvest, $\widehat{p}_N(\omega)$, in (13) into the true probability of the others observing NoInvest, F_N, where we denote

$$F_N \equiv F_B(k) + F_R(k). \tag{15}$$

Finally, we obtain simple *real population composition dynamics* such that

$$f(z_{t-1}) = \begin{cases} \beta z^2_{t-1} & \text{if } \Delta P_t(\omega, \ell_t) > 0, \\ -\beta z^2_{t-1} & \text{if } \Delta P_t(\omega, \ell_t) < 0, \end{cases} \tag{16}$$

[19] That is, from $P^B_{t-1}(\omega, \ell_{t-1})$ and $P^R_{t-1}(\omega, \ell_{t-1})$ in (11), by incorporating the true distributions $F_B(k)$ and $F_R(k)$ to $p_I(\omega^B)$ and $p_N(\omega^R)$, we have $\mathcal{P}^B_{t-1} = q_{t-2}(1 - F_B(k)) + (1 - q_{t-2})F_R(k)$ and $\mathcal{P}^R_{t-1} = (1 - q_{t-2})(1 - F_B(k)) + q_{t-2}F_R(k)$, so $A = 1 - F_B(k) + F_R(k)$.

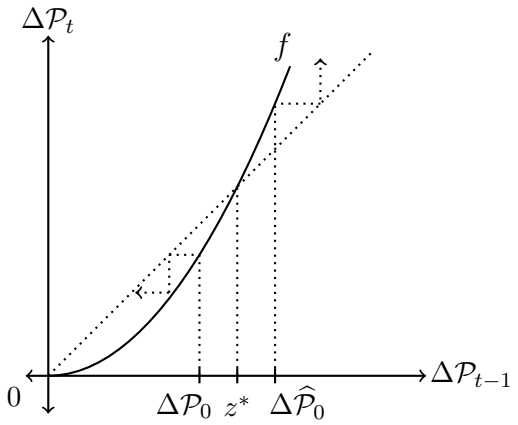

Figure 3: Real dynamics for $\Delta P_t(\omega, \ell_t) > 0$

where the coefficient β is to simplify the corresponding term, that is, $\beta \equiv \frac{1-F_N}{A(2-A)}$.[20] To appreciate the formula, one should find that it is *derived* only with the endogenous perception formation process and the law, without any exogenous functional form given. The dynamical system (16) is illustrated in Figure 3 with its fixed point z^* only for $\Delta P_t(\omega, \ell_t) > 0$.

The benchmark main result, Proposition 1, shows that if polarization is increasing (decreasing) in the current period, then it will continue to do so in the following period. This result reveals that for the binary splitting location equilibrium, there is no *interior absorbing* state for the dynamical system, essentially – as long as the initial population difference $\Delta\mathcal{P}_0$ is different from the fixed point z^*. The square function in (16) that characterizes the dynamical system thus embodies the vulnerability and risk that a society becomes polarized.

Proposition 1 *Suppose (A1)-(A5). Then, if $\widehat{p}_N(N, \emptyset) < 1$, the binary splitting location strategy is a location equilibrium. The associated two-period dynamical system satisfies the following properties.*

(i) If $|\Delta\mathcal{P}_t| > |\Delta\mathcal{P}_{t-1}|$, $|\Delta\mathcal{P}_{t+1}| > |\Delta\mathcal{P}_t|$.

(ii) If $|\Delta\mathcal{P}_t| < |\Delta\mathcal{P}_{t-1}|$, $|\Delta\mathcal{P}_{t+1}| < |\Delta\mathcal{P}_t|$.

For any arbitrary two periods in time, the equilibrium dynamics are either moving monotonically toward a perfectly mixed state or toward a perfectly polarized state. As we show in Sections 5 and 6, considering all possible histories and location strategies, the dynamics of our more general model inherit the same dynamical properties of Proposition 1.

Although the result appears to be derived seamlessly, it suffers from multiplicity of equilibria as follows. If players believe $\Delta P_t(\omega, \ell_t) > 0$ (resp. < 0), then the outcome *actually* arises from the binary splitting location strategy, when $1 - \widehat{p}_N(\omega) > 0$. Then, given (12), for

[20] Then, from the sum A_{t-1} and their difference $\Delta\mathcal{P}_{t-1}$, there exists a unique $(\mathcal{P}^B_{t-1}, \mathcal{P}^R_{t-1})$ such that $(\mathcal{P}^B_{t-1}, \mathcal{P}^R_{t-1}) = \left(\frac{A_{t-1}+\Delta\mathcal{P}_{t-1}}{2}, \frac{A_{t-1}-\Delta\mathcal{P}_{t-1}}{2}\right)$.

any q_{t-1} satisfying Lemma 1, both $\Delta P_t(\omega, \ell_t) > 0$ and $\Delta P_t(\omega, \ell_t) < 0$ for period t beliefs are possible. That is, regardless of whether there are actually more B members in the East in period $t-1$ or not, $\Delta \mathcal{P}_{t-1} > 0 \, (< 0)$, the equilibrium beliefs for the population difference in the next period t may arise in both ways: There can be more B or more R in the East in period t. Further, the equilibrium can change from one period to next, purely on the basis of beliefs; $\Delta P_{t-1}(\omega, \ell_{t-1}) > 0$ but $\Delta P_t(\omega, \ell_t) < 0$: There is a complete separation between the real part and the belief part of the system. However, since we have some proportion of *exogenous* location choices for the main model, *i.e.*, $\alpha \in (0, 1)$, we can construct a Markov location strategy to select a reasonable equilibrium in Section 5.

In addition, while the formula in (16) enlightens the key features of the dynamical system, several important properties remain unanswered which we address in the next sections. In particular, we need to show (i) the equilibrium payoffs of the homogeneous and the heterogeneous matches (Propositions 2, 3 and 4); (ii) the robust location strategy equilibrium (Proposition 5); (iii) the conditions for a complete mixed state, in particular, whether or not z^* is greater than 1 (Proposition 6); and (iv) the role of exogenous shocks to the system (Propositions 7 and 8). We now turn to providing such a complete equilibrium analysis; readers interested in the main polarization results can skip directly to Section 6 for Propositions 6-8.

4 Investment stage equilibrium

We begin with the equilibrium of the investment stage. Since we have analyzed the investment equilibrium of the young player (2) in Section 2, now we proceed to analyze the equilibrium of the old player. Unlike the young player, the equilibrium of the old player requires belief updating using Bayes' rule in (3). The analysis of the equilibrium for the old player can be further divided into homogeneous matches and heterogeneous matches. For both cases, given a symmetric equilibrium, we simplify the notation by omitting the subscripts i and j.

The friction in (A5) also makes our analysis more tractable in this section since it means that group members with all possible histories are always present with positive probability. That is, with this friction, players with all possible histories are on the equilibrium path so that we can still rely on a symmetric equilibrium.

4.1 Homogeneous match

In homogeneous matches (B, B) or (R, R), an old player is matched with a member of his own group. An old player's expected payoff from Invest depends primarily on his expectations about the probability that his matched partner chooses NoInvest, as for the young player's case in (1). However, in the old player's case, the NoInvest probability can differ according to different histories of interactions with own group members in youth. For instance, consider

a B player who is matched with a member of the B group when old. If the B player was also matched with his own group in youth and experienced Invest in the past, then the positive experience with a B group member makes him more optimistic about the distribution of his own group. Hence, the set of histories relevant for the homogeneous match case is given as $\Omega^S \equiv \{I, \emptyset, N\}$, where $\emptyset$ indicates no previous matching with the own group.

We apply in principle the same endogenous perception formation to homogeneous matches, but the process is relatively straightforward compared to heterogeneous matches. In order to derive the NoInvest probability, we start by denoting $X_t^S(\mathbf{k}_t^S)$ and $Y_t^S(\mathbf{k}_t^S)$ the probabilities that the partner chooses NoInvest conditional on F_X and NoInvest the probability conditional on F_Y, respectively. They depend on a threshold vector $\mathbf{k}_t^S$ that consists of the young player's k from (2) and the old player's history-contingent threshold profile $(k_t^S(\omega^S))_{\{\omega^S \in \Omega^S\}}$. Then, we can readily find them such that

$$\begin{aligned} X_t^S(\mathbf{k}_t^S) &= q_{t-1}(1 - F_X(k))F_X(k_t^S(I)) + q_{t-1}F_X(k)F_X(k_t^S(N)) + (1 - q_{t-1})F_X(k_t^S(\emptyset)), \\ Y_t^S(\mathbf{k}_t^S) &= q_{t-1}(1 - F_Y(k))F_Y(k_t^S(I)) + q_{t-1}F_Y(k)F_Y(k_t^S(N)) + (1 - q_{t-1})F_Y(k_t^S(\emptyset)). \end{aligned} \tag{17}$$

Each term of $X_t^S(\mathbf{k}_t^S)$ and $Y_t^S(\mathbf{k}_t^S)$ represents the probability of a partner choosing NoInvest given $r \in \{X, Y\}$, which can be derived with the partner's potential observations in youth and the history-contingent thresholds. For example, consider a homogeneous match (B, B). Then, a B player reasons that conditional on $r \in \{X, Y\}$, with probability q_{t-1} from (4), the matched B player previously met another B player when young, and with probability $(1 - F_r(k))$, he observed Invest from that youthful partner, where k is the equilibrium stage game threshold when young. For this observation, the corresponding threshold is given as $k_t^S(I)$. Together, this yields the first term in (17) of $X_t^S(\mathbf{k}_t^S)$ and $Y_t^S(\mathbf{k}_t^S)$, and will the other two terms accordingly. While the derivation is straightforward, note that for homogeneous matches, when a player forms perception about the other's perception, he assesses the partner's probability of observing any previous action *conditional* on the player's own beliefs about $r \in \{X, Y\}$ (*i.e.*, with probability $(1 - F_r(k))$). By contrast, for heterogeneous matches, even conditional on his own beliefs about the other group, the probability of the matched player from the other group observing an action from another player from his own group previously can be based on distribution being $r = X$ or $r = Y$.

Using $X_t^S(\mathbf{k}_t^S)$ and $Y_t^S(\mathbf{k}_t^S)$, we denote by $\Pr(N|\omega^S, \mathbf{k}_t^S)$ the overall probability of the partner choosing NoInvest for a homogeneous match, which is given as $\Pr(N|\omega^S, \mathbf{k}_t^S) = \pi(\omega^S)X_t^S(\mathbf{k}_t^S) + (1 - \pi(\omega^S))Y_t^S(\mathbf{k}_t^S)$. Hence, an old player with type θ and observation ω^S in $m = S$ obtains the expected payoff from Invest such that

$$U_t^S(\theta, \omega^S, \mathbf{k}_t^S) = d(\theta) - \Pr(N|\omega^S, \mathbf{k}_t^S), \tag{18}$$

whereas, as in the young player's case, NoInvest yields 0. Then, there exists a homogeneous match equilibrium if a history-contingent threshold profile $(k_t^S(\omega^S))_{\{\omega^S \in \Omega^S\}}$ makes the expected payoff from Invest in (18) equal to zero such that for each $\ell \in \{E, W\}$ and every $\omega^S \in \{I, \emptyset, N\}$,

$$d(k_t^S(\omega^S)) = \pi(\omega^S)X_t^S(\mathbf{k}_t^S) + (1 - \pi(\omega^S))Y_t^S(\mathbf{k}_t^S). \tag{19}$$

The history-contingent threshold profile, unlike the equilibrium for the young in (2), makes the old player's equilibrium a fixed point $(k_t^S(\omega^S))_{\{\omega^S \in \Omega^S\}}$ of a *multivariable* mapping. More interestingly, a homogeneous match results in the following monotonicity relationship. For each pair $\omega^S, \widehat{\omega}^S$ satisfying $\pi(\widehat{\omega}^S) > \pi(\omega^S)$, we have $k_t(\widehat{\omega}^S) < k_t(\omega^S)$.[21]

Proposition 2 *Suppose (A1)-(A5). Then for each $t = 1, 2, ...,$ a homogeneous match equilibrium profile of the old player's thresholds satisfies monotonicity in that*

$$k_t^S(I) < k_t^S(\emptyset) < k_t^S(N).$$

This result is quite intuitive. If a B member observes Invest from another B member when they were young, he is more likely to invest in the current (B, B) match because the positive experience makes him more optimistic about group B and also because he thinks that his partner is more likely to invest as well from having the same good experience. This yields the first inequality – a lower stage game threshold means a higher probability of choosing Invest – and the same argument works for the second inequality.

The history-contingent equilibrium of homogeneous matches among old members is interesting, but the history of play in interactions with members of *the other* group in youth is not relevant in this case. This will no longer hold in the case of heterogeneous matches, as shown in the next subsection.

4.2 Heterogeneous match

In heterogeneous matches (B, R), an old player is matched with a member of the other group. As in the young player's payoff from Invest in (1) and the homogeneous payoff in (18), the old player's beliefs about the probability that his matched partner will choose NoInvest play an essential role. We denote that probability in the heterogeneous match case by $\Pr(N|\omega^A, \mathbf{k}_t^A)$, where $\omega^A \in \Omega^A$ is an old player's history for a heterogeneous match, and a threshold vector $\mathbf{k}_t^A$ consists of the young's k and the old's history-contingent threshold profile $(k_t^A(\omega^A))_{\{\omega^A \in \Omega^A\}}$. However, in this case, the probability of the partner choosing NoInvest depends both on how good his own group is and how good the matched partner's group is, apart from the partner's strategy of choosing NoInvest given each history. That is, unlike in the homogeneous match case, histories from both the same and the other group can matter in the heterogeneous match case since each player cares not only about how good the other group is but also how the matched partner of the other group evaluates *his own* group.

The subtle role of the histories of play for the behavior of agents in old age in the heterogeneous matches can be illustrated as follows. Consider a B player who is matched with a member of group R when old. If the B player was also matched with a member of

[21] Note that to simplify the notation in both the homogeneous match and a heterogeneous match below, we omit q_{t-1}, but all functions depend on q_{t-1}, the previous "stock" – as well as other parameters like F_X and F_Y – so the value of $k_t(\omega^S)$ changes as q_{t-1} changes. Nonetheless, the monotonicity holds for any $q_{t-1} > 0$. The role of q_{t-1} becomes apparent in Proposition 4 and especially in the dynamics found in Section 6.

group R in youth and experienced Invest, this increases the likelihood that the B player will invest in the old age match with the R member since the B player is more optimistic about the distribution of group R, extending the argument in the homogeneous match case. Suppose, instead, that the B player was matched with a member of his own B group in youth and experienced the Invest outcome. Then, the B player thinks it more likely that his current counterpart, the old R member, may also have had a good experience with group B in youth, which makes the B player more likely to choose Invest in the old age match with the R member. That is, the second case also increases the likelihood that the B player will invest in the old age match with the R member. As we show below in Proposition 3, the former effect is stronger than the latter effect so that a B player is more likely to invest in the old age match with an R member if he observed a member of the R group investing in the past than if he observed a member of his own B group investing in the past (see the first inequality of Proposition 3). To analyze it more formally, let an old player's history for a heterogeneous match be denoted by $\omega^A \in \Omega^A \equiv \{I|\emptyset, \emptyset|I, \emptyset|N, N|\emptyset\}$. The first observation is one from the matched partner's group (different group), and the second observation is from the old player's own group, when the old player was young; for example, from a B player's perspective, $\emptyset|I$ indicates that, previously, the B member was not matched with an R member, but instead observed Invest from another B member.

As discussed, the perception formation process in this case differs from that of homogeneous matches. In contrast to the homogeneous case in (18), even conditional on the matched R group distribution being $r = X$ or $r = Y$, the uncertainty with respect to the player's own B group distribution remains. This uncertainty can be captured by those probabilities $p_I(\omega^B)$ and $p_N(\omega^B)$ from (5). That is, $p_I(\omega^B)$ (resp. $p_N(\omega^B)$) is an old B member's belief that his matched R player previously observed Invest (resp. NoInvest) from another B player. With these probabilities, we can then derive the probability of a matched partner choosing NoInvest as $\Pr(N|\omega^A, \mathbf{k}_t^A) = \pi(\omega^A)X_t^A(\mathbf{k}_t^A, \omega^B) + (1 - \pi(\omega^A))Y_t^A(\mathbf{k}_t^A, \omega^B)$ in a (B, R) match. Similar to $X_t^S(\mathbf{k}_t^S)$ and $Y_t^S(\mathbf{k}_t^S)$ in the homogeneous match case, each of $X_t^A(\mathbf{k}_t^A, \omega^B)$ and $Y_t^A(\mathbf{k}_t^A, \omega^B)$ in this case denotes the probability of a partner choosing NoInvest conditional on $r \in \{X, Y\}$, but unlike them, $X_t^A(\mathbf{k}_t^A, \omega^B)$ and $Y_t^A(\mathbf{k}_t^A, \omega^B)$ depend on ω^B through $p_I(\omega^B)$ and $p_N(\omega^B)$, given the subtlety illustrated above.

Hence, an old player with type θ and observation ω^A in $m = A$ obtains an expected payoff from Invest such that

$$U_t^A(\theta, \omega^A, \mathbf{k}_t^A) = d(\theta) - \Pr(N|\omega^A, \mathbf{k}_t^A). \tag{20}$$

Then, there exists a heterogeneous match equilibrium if a history-contingent threshold profile $(k_t^A(\omega^A))_{\{\omega^A \in \Omega^A\}}$ makes the expected payoff from Invest in (20) equal to zero such that for each $\ell \in \{E, W\}$ and every $\omega^A \in \Omega^A$,

$$d(k_t^A(\omega^A)) = \pi(\omega^R)X_t^A(\mathbf{k}_t^A, \omega^B) + (1 - \pi(\omega^R))Y_t^A(\mathbf{k}_t^A, \omega^B). \tag{21}$$

As in the homogeneous match case, the old player's equilibrium is a fixed point of a multivariable mapping but now one that has four dimensions. Further, one has to interpret

equation (21) carefully. On the left-hand side (LHS), $k_t^A(\omega^A)$ is from an old B player's point of view, that is, $\omega^A = \omega^R|\omega^B$, which is incorporated into $\pi(\omega^R)$ and $\pi(\omega^B)$ through $p_I(\omega^B)$ and $p_N(\omega^B)$ in $X_t^A(\mathbf{k}_t^A, \omega^B)$ and $Y_t^A(\mathbf{k}_t^A, \omega^B)$ above. On the right-hand side (RHS), $\mathbf{k}_t^A$ is a vector of thresholds taken by the matched old R player, so a history inside any such threshold is interpreted as $\omega^B|\omega^R$, *i.e.*, from an R member's perspective.[22]

Since $X_t^A(\mathbf{k}_t^A, \omega^B)$ and $Y_t^A(\mathbf{k}_t^A, \omega^B)$ depend on ω^B, finding a monotonicity result for a profile of the old player's thresholds k_t^A is not straightforward. While, to some degree, the monotonicity between $k_t^A(\emptyset|I)$ and $k_t^A(\emptyset|N)$ and that between $k_t^A(I|\emptyset)$ and $k_t^A(N|\emptyset)$ resemble those found in the homogeneous case in Proposition 2, the monotonicity between $k_t^A(I|\emptyset)$ and $k_t^A(\emptyset|I)$ or that between $k_t^A(N|\emptyset)$ and $k_t^A(\emptyset|N)$ requires a whole new approach.

Proposition 3 *Suppose (A1)-(A5). Then, for each $t = 1, 2, ...,$ a heterogeneous match equilibrium profile of the old player's thresholds satisfies monotonicity in that*

$$k_t^A(I|\emptyset) < k_t^A(\emptyset|I) < k_t^A(\emptyset|N) < k_t^A(N|\emptyset).$$

Hence, in a heterogeneous match (B, R), as discussed earlier, the relationship $k_t^A(I|\emptyset) < k_t^A(\emptyset|I)$ means that a B player is more likely to invest if he observed Invest from an R player than if he observed Invest from another B player when young. On the other hand, the relationship $k_t^A(\emptyset|N) < k_t^A(N|\emptyset)$ means that a B player is more likely to invest if he observed NoInvest from another B player than if he observed NoInvest from an R player when young. In sum, a B player's experience in the past with a young R player *reinforces* his investment decision with a matched old R player in either direction, compared with the same experience with a young B player. Interestingly, even if the B player has no history in youth with an R player, we have $k_t^A(\emptyset|I) < k_t^A(\emptyset|N)$. The intuition is that if the old B player met a B player in youth who chose to Invest (resp. NoInvest), then the old B player believes that his current matched R partner is more likely to have also had a good (resp. bad) experience with a B player when he was young, making the old B player more (resp. less) likely to choose Invest in the match with the old R player.

5 Location stage equilibrium

In this section, we determine a dynamic spatial equilibrium, which consists of three components: the young player's investment stage equilibrium, the old player's investment stage equilibrium, and a location stage equilibrium. With the equilibrium of the young and that of the old players in previous sections, finding a dynamic spatial equilibrium boils down to characterizing a location stage equilibrium. In doing so, we allow for all feasible location strategies, while previously in Section 3, we consider a particular location strategy, the binary splitting location strategy.

[22]This is precisely why we need a notation different from the neutral notation $\omega = (\omega^B, \omega^R)$ in Section 3.

First, we find the sets of histories that yield favorable and unfavorable stances toward one's own group, using the previous investment stage equilibrium characterizations from homogeneous and heterogeneous matches. To compare the two future payoffs when old, one from homogeneous matches in (18) and another from heterogeneous matches in (20), we reconsider the neutral notation $\omega = (\omega^B, \omega^R)$ from a B player's perspective, as in Section 3, and also a B player's expected payoff for a location strategy, without loss of generality. If a B player chooses E, he obtains the expected payoff

$$\begin{aligned} V_t^B(E, \theta, \omega, \ell_t) = &\frac{P_t^B(\omega, \ell_t^B)}{P_t^B(\omega, \ell_t^B) + P_t^R(\omega, \ell_t^R)} U_t^S(\theta, \omega^S, \mathbf{k}_t^S) \\ &+ \frac{P_t^R(\omega, \ell_t^R)}{P_t^B(\omega, \ell_t^B) + P_t^R(\omega, \ell_t^R)} U_t^A(\theta, \omega^A, \mathbf{k}_t^A), \end{aligned}$$

where $P_t^B(\omega, \ell_t^B)$ and $P_t^R(\omega, \ell_t^R)$ are from (7) together with $\ell_t \equiv (\ell_t^B, \ell_t^R)$ in (6), whereas $U_t^S(\theta, \omega^S, \mathbf{k}_t^S)$ and $U_t^A(\theta, \omega^A, \mathbf{k}_t^A)$ are from (18) and (20) in the old's investment stage. If, on the other hand, the B player chooses W, he obtains the expected payoff

$$\begin{aligned} V_t^B(W, \theta, \omega, \ell_t) = &\frac{1 - P_t^B(\omega, \ell_t^B)}{2 - P_t^B(\omega, \ell_t^B) - P_t^R(\omega, \ell_t^R)} U_t^S(\theta, \omega^S, \mathbf{k}_t^S) \\ &+ \frac{1 - P_t^R(\omega, \ell_t^R)}{2 - P_t^B(\omega, \ell_t^B) - P_t^R(\omega, \ell_t^R)} U_t^A(\theta, \omega^A, \mathbf{k}_t^A). \end{aligned}$$

With these expected payoffs, we can now provide a formal definition for a dynamic spatial equilibrium at each period $t = 1, 2, ...$

Definition 1 $\left(k, (k_t^S(\omega^S))_{\{\omega^S \in \Omega^S\}}, (k_t^A(\omega^A))_{\{\omega^A \in \Omega^A\}}, \left(\ell_t^g\right)_{g \in \{B,R\}}\right)$ *constitutes a dynamic spatial equilibrium at* $t = 1, 2, ...$ *if for each* $g \in \{B, R\}$*,*

(i) Each young player's threshold k *satisfies* $U(k, k) = 0$ *in (2).*

(ii) For homogeneous matches, each old player's threshold $k_t^S(\omega^S)$ *for every* $\omega^S \in \Omega^S$ *satisfies* $U_t^S(k_t^S(\omega^S), \omega^S, \mathbf{k}_t^S) = 0$ *in (19), and for heterogenous matches, each old player's threshold* $k_t^A(\omega^A)$ *for every* $\omega^A \in \Omega^A$ *satisfies* $U_t^A(k_t^A(\omega^A), \omega^A, \mathbf{k}_t^A) = 0$ *in (21).*

(iii) For each $(\theta, \omega) \in \Theta \times \Omega$*,* $V_t^g(\ell_t^g(\theta, \omega, z_{t-1}), \theta, \omega, \ell_t) \geq V_t^g(\ell, \theta, \omega, \ell_t)$ *for all* $\ell \in \{E, W\}$*.*

(iv) For each $\omega \in \Omega$*,* $P_t^g(\omega, \ell_t^g) = \alpha \mathbb{E}\left[\mathbf{1}_{\{\ell_t^g(\widetilde{\theta}, \widetilde{\omega}, z_{t-1}) = E\}} \mid \omega\right] + (1 - \alpha)\mathcal{P}_{t-1}^g$*.*

Although the definition is general, given the investment stage equilibrium of the young in Section 2 and that of the old in Section 4, a dynamic spatial equilibrium at each period t is reduced to a location (Bayesian) equilibrium satisfying (iii) and (iv). The first condition (iii) of a location equilibrium addresses optimality, and the second condition (iv) addresses consistency such that each player's expectations about the other players' location strategies are correct. For example (but repeating the procedure in Section 3), consider a player with $\omega = (I, \emptyset)$ who forms beliefs about $P_t^B(\omega, \ell_t^B)$ in expectation of $\ell_t^B(\widetilde{\theta}, \widetilde{\omega}, z_{t-1})$ for all four

possible histories $\widetilde{\omega} \in \Omega$, where $\widetilde{\omega}$ denotes a history that the other player can have, and in equilibrium, the expectations must be "rational" in the sense that they are identical to the actual equilibrium choices of players with other histories. However, this does *not* mean that $P_t^B(\omega, \ell_t^B) = \mathcal{P}_t^B$, that is, rational expectations are not stretched to the degree that players are capable of expecting the exact number of B members in E based on the *true* distributions.

We first show that there is no role for a player's intrinsic type θ in the location stage decision; the only private information that matters for the location equilibrium is the *history* of observations ω.

Lemma 2 *Suppose (A1)-(A5). Then, the payoff difference between a homogeneous match and a heterogeneous match is equivalent to the difference in their corresponding thresholds such that* $U_t^S(\theta, \omega^S, \mathbf{k}_t^S) - U_t^A(\theta, \omega^A, \mathbf{k}_t^A) = d(k_t^A(\omega^A)) - d(k_t^S(\omega^S))$.

Then, the location decision (6) made by a player who is a member of group $g \in \{B, R\}$ and who resides in either location becomes a function only of ω and a previous real population difference z_{t-1} such that $\ell_t^g : \Omega \times [-1, 1] \to \{E, W\}$. The lemma also provides an important intermediate step for how one can actually find an equilibrium using the location equilibrium definition in Definition 1. In equilibrium, the optimal location decisions are given as follows.

$$\begin{aligned} &\text{Any } B \text{ member with } \omega \text{ chooses } E \text{ if } \Delta P_t(\omega, \ell_t)\left[d(k_t^A(\omega^A)) - d(k_t^S(\omega^S))\right] > 0. \\ &\text{Any } R \text{ member with } \omega \text{ chooses } E \text{ if } -\Delta P_t(\omega, \ell_t)\left[d(k_t^A(\omega^A)) - d(k_t^S(\omega^S))\right] > 0. \end{aligned} \tag{22}$$

The intuition behind these equilibrium location decisions initially seems straightforward, as players tend to move toward the location where they anticipate higher payoffs. However, these decision rules also reveal the intricate nature of the problem, the endogenous perception formation process. To see that more clearly, let us delve further into the conditions in (22). These conditions can be divided into two components: $d(k_t^A(\omega^A)) - d(k_t^S(\omega^S))$ and $\Delta P_t(\omega, \ell_t)$. Now, consider a particular history $\omega = (\omega^B, \omega^R)$ among the four histories in Ω. Then, we need to determine whether ω, by incorporating $\omega = (\omega^B, \omega^R)$ into ω^S and ω^A, yields $d(k_t^A(\omega^A)) > d(k_t^S(\omega^S))$ or not. In other words, the first critical element is to *classify* what histories make each player expect such a favorable stance toward their own group. This, however, is not sufficient for the location equilibrium analysis, since a player with a history favorable toward his own group wants to anticipate "correctly" which location has more of the same group members. Note that the belief difference $\Delta P_t(\omega, \ell_t)$ contains location strategies for *all* four histories, as already discussed in Section 3. This implies that a location equilibrium exists only when, for each history $\omega \in \Omega$, a player with ω selects a location that is "compatible" with their incentive to choose that location, given the player's rational expectations about $\Delta P_t(\omega, \ell_t)$ that all other players with different histories choose locations exactly as expected – the way we use rational expectations is essentially no different from that of the standard Nash equilibrium or general equilibrium. Since all players simultaneously choose their location strategies, in a location stage equilibrium, the

beliefs held by the two group members, across all four possible histories, must "clear" so as to satisfy the consistency condition (iv) of Definition 1.

To tackle the first component (iii) of a location equilibrium, we start by defining a B player's set of histories that yield favorable and unfavorable stances toward his own group as Ω^{B+} and Ω^{B-}, which are equal to a set of histories satisfying $U_t^S(\theta, \omega^S, \mathbf{k}_t^S) > U_t^A(\theta, \omega^S, \mathbf{k}_t^A)$ and a set of histories satisfying $U_t^S(\theta, \omega^S, \mathbf{k}_t^S) < U_t^A(\theta, \omega^S, \mathbf{k}_t^A)$, respectively. By Lemma 2, they can be defined as

$$\Omega^{B+} \equiv \left\{\omega \in \Omega : k_t^A(\omega^A) > k_t^S(\omega^S)\right\} \text{ and } \Omega^{B-} \equiv \left\{\omega \in \Omega : k_t^A(\omega^A) < k_t^S(\omega^S)\right\}, \tag{23}$$

where note that a lower threshold means a higher probability of choosing Invest. The corresponding sets can be defined for R players as well. Now, to identify the sets Ω^{g+} and Ω^{g-} reduces to comparing a profile of the old player's thresholds from a homogeneous match equilibrium in Proposition 2 with that from a heterogeneous match equilibrium in Proposition 3 in the previous section. The two profiles of thresholds are fixed points from two mappings, which in turn implies that the comparison must be made between the two fixed points and thus between the two mappings – the three-dimensional homogeneous mapping and the four-dimensional heterogeneous mapping.

For the comparison, we need an additional condition for monotonicity of the two mappings as in Milgrom and Shannon (1994). Since the monotonicity is not germane to the model, we provide the assumption now:

(A6) For $m \equiv \max\left\{\frac{F_Y(k)}{F_X(k)}, \frac{1-F_X(k)}{1-F_Y(k)}\right\}$, $F_X(\theta') - F_X(\theta) \geq m\big[F_Y(\theta') - F_Y(\theta)\big]$ for all $\theta, \theta' \in \Gamma$.

Assumption (A6) is satisfied for a large class of distributions. For example, a sufficient condition for (A6) is $\frac{f_X(\theta)}{f_Y(\theta)} \geq m$ for all $\theta \in \Gamma$, where $\frac{f_X(\theta)}{f_Y(\theta)}$ is the likelihood ratio, which can be called the *bounded* likelihood ratio condition (a weaker version of (A6) can be found in Section 7). In addition to the parameterized monotone comparative statics analysis, this paper also departs from Milgrom and Shannon (1994) in that we need to determine not only whether a fixed point increases or not but also by *how much* it changes for the comparison, as we discuss subsequently.

We establish the monotone comparison between two sets of thresholds for the symmetric and asymmetric match cases.[23]

Proposition 4 *(Comparison between two types of matches) Suppose (A1)-(A6). Then, there exists $\Delta\overline{\mathcal{P}}_0$ such that for each initial population difference $\Delta\mathcal{P}_0 \geq \Delta\overline{\mathcal{P}}_0$, in equilibrium at $t = 1, 2....$, the relationship between the thresholds in a homogeneous match and the thresholds in a heterogeneous match is given as follows.*

[23]There can be multiple fixed points since (A3) is not strong enough to guarantee uniqueness for k_t^S. In that case, following the typical treatment of the standard monotone comparative statics analysis approach, we suppose that an equilibrium arises either at the largest point or at the smallest point. A condition slightly stronger than (A3) can guarantee uniqueness for k_t^S, which is interestingly related to m in (A6) as well. We discuss them in Section 7.

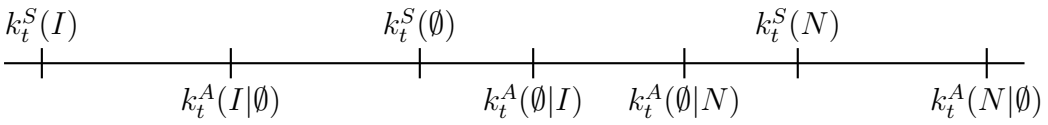

Figure 4: Comparison between two matches

(i) $k_t^S(I) < k_t^A(\emptyset|I)$ *and* $k_t^S(\emptyset) < k_t^A(N|\emptyset)$.

(ii) $k_t^S(N) > k_t^A(\emptyset|N)$ *and* $k_t^S(\emptyset) > k_t^A(I|\emptyset)$.

We obtain the first set of results in (i) of Proposition 4 from the monotone comparative statics, but the second set of results in (ii) above requires that the heterogeneous equilibrium thresholds *do not* increase too much (see Appendix for the formal proof). In particular, regarding (ii), it can be readily shown that as the society approaches the perfectly polarized state, that is, for q_{t-1} close to 1, the two equilibrium thresholds, the fixed point in Proposition 2 and the one in Proposition 3, are sufficiently close to each other that they satisfy the second set of results.[24]

Summarizing the discussion to this point, Figure 4 shows the relationships between homogeneous and heterogeneous matches from Proposition 4 that arise out of Propositions 1-3. The resulting set of histories favorable and unfavorable to playing the game with own group members yields the history sets in (10) from Section 3. The four comparisons reveal how the same histories can lead to a difference in future expected payoffs when a player is matched with a member of his own group and when he is matched with a member of the other group. That is, from a B player's point of view, if he met another B player when young and observed Invest (resp. NoInvest), *i.e.*, $\omega = (I, \emptyset)$ (resp. $\omega = (N, \emptyset)$), a match with a member of the same B group when old yields a higher (resp. lower) future expected payoff than does a match with an R player, that is, $k_t^S(I) < k_t^A(\emptyset|I)$ (resp. $k_t^S(N) > k_t^A(\emptyset|N)$) – a lower threshold means a higher expected payoff as shown in (22). Overall, a good (bad) experience with a member of the same group enlarges (diminishes) the future payoff from matching with another member of the same group, and a similar reasoning applies to the cases of matching with a member of the other group in the past, *i.e.*, $\omega = (\emptyset, I)$ or $\omega = (\emptyset, N)$.[25]

Now, to verify the consistency part (iv) of Definition 1, we proceed in the following two steps. Lemma 2 simplifies a location strategy so that it is a function of a history ω and a

[24]Later on in the paper, to establish a polarization convergence result in Proposition 7, we will need another critical value for the initial population difference that is separate from the critical value referenced in Proposition 4. In the proof of Proposition 7, we will use the maximum of these two critical values, which will ensure that the latter proposition continues to hold.

[25]Despite the equality $\Omega^{B+} = \Omega^{R-}$, one needs to be careful about interpretations since $(I, \emptyset)$ in Ω^{B+} means that a B player observed Invest from a member of the same group, whereas $(I, \emptyset)$ in Ω^{R-} means that an R player observed Invest from a member of a different group, a B group member. Hence, in terms of its content, the set Ω^{R-} is identical to Ω^{B-}.

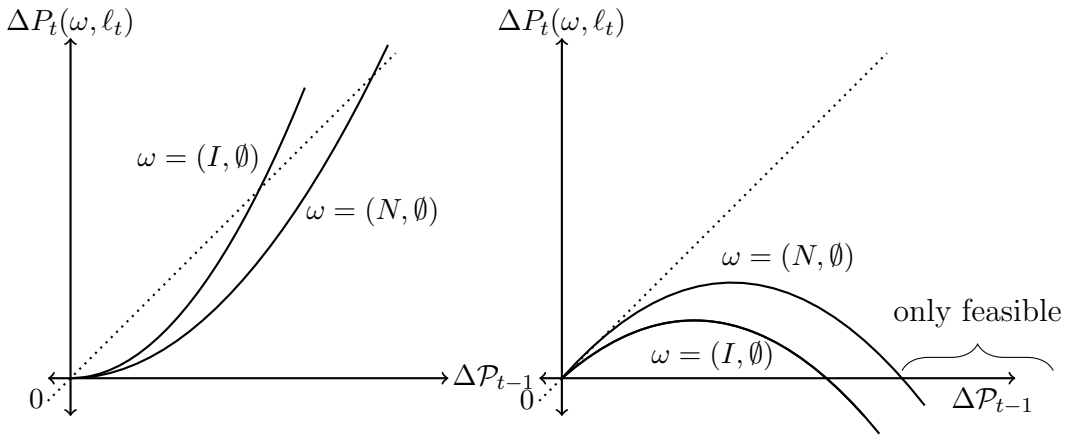

Figure 5: Belief dynamics for $\Delta P_t(\omega, \ell_t) > 0$ (left) and for $\Delta P_t(\omega, \ell_t) < 0$ (right)

previous population difference $\Delta \mathcal{P}_{t-1}$ in the two locations. In the first step, we consider the case $\alpha = 1$, and next, we allow $\alpha \in (0, 1)$ with a concept of robustness of a location strategy, which also clarifies the role of robustness. Suppose $\alpha = 1$, and observe that there are 2^8 strategies to consider given that each group member has four different histories Ω^{g+}, Ω^{g-} in (10) with two location choices in (2). Despite the large number of strategies, the following lemma shows that we can significantly reduce the number of possible strategies for a location equilibrium.

Lemma 3 *Suppose (A1)-(A6) and $\alpha = 1$. Then for $\Delta P_t(\omega, \ell_t) \neq 0$, a location equilibrium ℓ_t satisfies the following properties.*

(i) For each $\omega \in \Omega$, $\ell_t^B(\omega) = \ell_t^R(\omega)$.

(ii) For each $g \in \{B, R\}$, $\ell_t^g(\omega) = \ell_t^g(\omega')$ for all $\omega \neq \omega' \in \Omega^{g+}$ and $\ell_t^g(\omega) = \ell_t^g(\omega')$ for all $\omega \neq \omega' \in \Omega^{g-}$.

Then, by Lemma 3, for $\alpha = 1$, if there exists a location equilibrium, then by its very nature, it will be either the binary splitting strategy given in Table 2 of Section 3 or a strategy with all players moving to one location. Proposition 5 shows that those two properties of the lemma continue to hold for $\alpha \in (0, 1)$. Now, if $\alpha \in (0, 1)$, the latter strategy can no longer be an equilibrium for $\Delta \mathcal{P}_{t-1} \neq 0$.[26] Then, to examine whether the binary splitting strategy survives under $\alpha \in (0, 1)$, by extending the belief dynamical system in (14), we obtain

$$\Delta P_t(\omega, \ell_t) = \begin{cases} \alpha \frac{1 - \widehat{p}_N(\omega)}{A_{t-1}(2 - A_{t-1})} \Delta \mathcal{P}_{t-1}^2 + (1 - \alpha) \Delta \mathcal{P}_{t-1} & \text{if } \Delta P_t(\omega, \ell_t) > 0, \\ -\alpha \frac{1 - \widehat{p}_N(\omega)}{A_{t-1}(2 - A_{t-1})} \Delta \mathcal{P}_{t-1}^2 + (1 - \alpha) \Delta \mathcal{P}_{t-1} & \text{if } \Delta P_t(\omega, \ell_t) < 0, \end{cases} \tag{24}$$

[26] A strategy with all players moving to one location – precisely, for each $\omega \in \Omega^{g+}$, $\omega' \in \Omega^{g-}$, $\ell_t^g(\omega) = \ell_t^g(\omega')$ – requires $\Delta P_t(\omega, \ell_t) = 0$. However, the strategy leads to $\mathbb{E}\big[\mathbf{1}_{\{\ell_t^B(\widetilde{\omega}, z_{t-1}) = E\}} \mid \omega\big] - \mathbb{E}\big[\mathbf{1}_{\{\ell_t^R(\widetilde{\omega}, z_{t-1}) = E\}} \mid \omega\big] = 0$, which yields $\Delta P_t(\omega, \ell_t) = (1 - \alpha) \Delta \mathcal{P}_{t-1} \neq 0$, so we have a contradiction. Also, this together with (ii) of Lemma 3 implies that players with any two different histories $\omega \neq \omega'$ share the same beliefs about the sign of $\Delta P_t(\omega, \ell_t)$ such as either $\Delta P_t(\omega, \ell_t) > 0, \Delta P_t(\omega', \ell_t) > 0$ or $\Delta P_t(\omega, \ell_t) < 0, \Delta P_t(\omega', \ell_t) < 0$.

where, unlike A in (14), now the sum of the two moving probabilities $A_{t-1} \equiv \mathcal{P}^B_{t-1} + \mathcal{P}^R_{t-1}$ must be denoted with a time subscript. Note that the more frequent matching with members of the same group in Lemma 1 generates a *strictly convex* shape for the belief dynamics in (24), as illustrated in Figure 5.[27]

In Section 3, we show that for the two-period dynamical system in (14), only with strategic location choices, a previous population composition has no "control" of the next period beliefs at all; that is, for any population difference $\Delta\mathcal{P}_{t-1} > 0$, both belief differences $\Delta P_t(\omega, \ell_t) > 0$ and $\Delta P_t(\omega, \ell_t) < 0$ can arise. With the addition of an exogenous force α, and using a Markov strategy, we can relate the real part $\Delta\mathcal{P}_{t-1}$ and the belief part $\Delta P_t(\omega, \ell_t)$ in a reasonable way. That is, given the belief dynamics (24), for $\Delta\mathcal{P}_{t-1} > 0$, we could have $\Delta P_t(\omega, \ell_t) < 0$ (the right figure of Figure 5), as well as $\Delta P_t(\omega, \ell_t) > 0$, but if the negative sign arises, it does so only when $\Delta\mathcal{P}_{t-1} > 0$ is sufficiently large, with the horizontal intercept $\frac{(1-\alpha)\gamma}{\alpha}\frac{A_{t-1}(2-A_{t-1})}{1-\widehat{p}_N(\omega)}$. This means that if there are sufficiently more B members in E, then people somehow anticipate that, subsequently, there is such a *shift* in the population composition, so that there are more B members in W next period. However, by adopting a robust location strategy in Section 2, we erase such unreasonable, non-robust outcomes. Note that a location strategy with $\Delta\mathcal{P}_{t-1} > 0$ and $\Delta P_t(\omega, \ell_t) < 0$ is not robust because it is valid for a particular size $\alpha \in (0, 1)$.

Consider a *Markov binary splitting location strategy* ℓ_t, which states that: for each $\omega \in \Omega^{B+} = \Omega^{R-}$, $\omega' \in \Omega^{B-} = \Omega^{R+}$,

$$\begin{cases} \ell^g_t(\omega, z_{t-1}) = E, \ell^g_t(\omega', z_{t-1}) = W & \text{if } z_{t-1} \geq 0, \\ \ell^g_t(\omega, z_{t-1}) = W, \ell^g_t(\omega', z_{t-1}) = E & \text{if } z_{t-1} \leq 0. \end{cases} \tag{25}$$

We can identify the necessary and sufficient condition for the existence of a location equilibrium and show the above Markov strategy in (25) is only one that generates a robust location equilibrium, which satisfies $\Delta P_t(\omega, \ell_t) = 0$ if and only if $\Delta\mathcal{P}_{t-1} = 0$.

Proposition 5 *Suppose (A1)-(A6). Then, for each $t = 1, 2, 3, ...$, there exists a robust location equilibrium strategy if and only if $\widehat{p}_N(N, \emptyset) < 1$; and further, any robust location strategy is a Markov location strategy in (25).*

By focusing on a Markov strategy, there is no possibility of any oscillation in the belief dynamics; that is, for $\Delta\mathcal{P}_{t-1} > 0$, we only have belief dynamics as in the left panel of Figure 5. However, this does not mean that we are excluding the possibility of oscillation in terms of the *real* dynamics, as will be analyzed in detail next section.

[27]Note that for Figures 5, in the case of $\Delta P_t(\omega, \ell_t) < 0$, by symmetry, the negative dimension's graph looks exactly the same as the reversed one in the positive dimension in the case of $\Delta P_t(\omega, \ell_t) > 0$.

6 Matching dynamics and polarization

In this section, we study the matching dynamics of the system over time. We do so from the perspective of an outside theorist who perfectly knows the distributions F_B and F_R.

In the model, recall (A5) that some fraction of the population do not adjust their location. We now add an amplifying effect ϵ to this external force so that we can study how likely we are to observe polarization as the limiting outcome of our system.[28] As we show, the extent of polarization will depend on the size of this amplifying effect. Specifically, consider the effect for group B and suppose there are *more* B members in the East, *i.e.*, $\Delta\mathcal{P}_{t-1} > 0$. Under the amplifying effect for some $\epsilon > 0$, there will be additional $\epsilon\Delta\mathcal{P}_{t-1}$ proportion of B members who *do not change* their location since there are more B members in the East. Symmetrically, if there are *fewer* B members in the East, *i.e.*, if $\Delta\mathcal{P}_{t-1} < 0$, then an additional $\epsilon\Delta\mathcal{P}_{t-1}$ proportion of B members *change* their location to the West, where there are more B members. Hence, for proportion $1-\alpha$, $\mathcal{P}^B_{t-1} + \epsilon\Delta\mathcal{P}_{t-1}$ can be interpreted as the B group's *inertia coupled with an amplifying effect* given the difference in the population size of the two groups $\Delta\mathcal{P}_{t-1}$ in the East. An identical external force is applied to R group members. Let $\gamma \equiv 1 + 2\epsilon > 1$ denote the sum of the inertia and the gross amplifying effect since the first term 1 is from pure inertia and the second term, 2ϵ, comes from the combined amplifying effects of both groups.

We incorporate the Markov location strategy in (25) given Proposition 5 into the actual dynamics with the two true distributions, F_B and F_R. In other words, we turn the belief dynamics in (24) into real dynamics by replacing $\widehat{p}_N(\omega)$ in (13) by F_N in (15), as in Section 3. With systematic polarization $\gamma > 1$, a system of the real population composition dynamics is given by $z_t = f_{t-1}\big(z_{t-1}\big)$ for $t = 1, 2, ...$ such that

$$f_{t-1}(z_{t-1}) \equiv \begin{cases} \alpha\beta_{t-1}z^2_{t-1} + (1-\alpha)\gamma z_{t-1} & \text{if } z_{t-1} \geq 0, \\ -\alpha\beta_{t-1}z^2_{t-1} + (1-\alpha)\gamma z_{t-1} & \text{if } z_{t-1} \leq 0, \end{cases} \tag{26}$$

where we denote

$$\beta_{t-1} \equiv \frac{1 - F_N}{A_{t-1}(2 - A_{t-1})}. \tag{27}$$

Despite the complex procedure to characterize a location equilibrium in the previous section, the resulting dynamical system is relatively simple. The figure for the system is Figure 6, which is similar to Figure 3 from Section 3. The system in (26) is a version of the *square function* in (16) found earlier in Section 3 in the case where $\alpha = 1$. The key difference between these two systems is how its dynamical system operates using the sum of the two moving probabilities. With $\alpha < 1$, $A_{t-1} \equiv \mathcal{P}^B_{t-1} + \mathcal{P}^R_{t-1}$ changes over time (see Lemma 4) and thereby, we have a time subscript f_{t-1} through β_{t-1} in (27) in the system. For the extended

[28] We do not think of this effect ϵ as an assumption, but rather as an exogenous parameter. In this section, we show how our polarization result hinges on the size of this friction which plays a role in the comparative statics analysis.

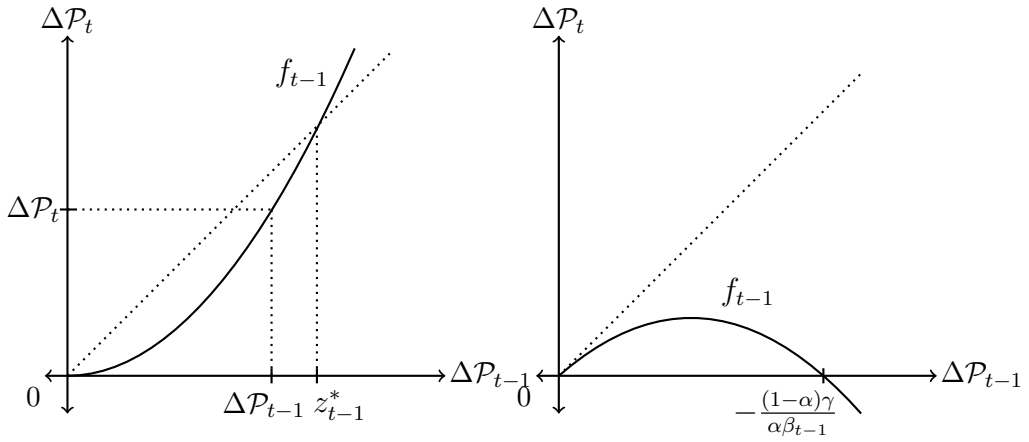

Figure 6: Real dynamics for mixed state if $1 - F_N > 0$ (left) and $1 - F_N < 0$ (right)

system with both endogenous and exogenous location choices in (26), we denote each fixed point of f_{t-1} by z^*_{t-1}, *i.e.*, $z^*_{t-1} = f_{t-1}(z^*_{t-1})$, where the fixed point is given by:

$$z^*_{t-1} = \frac{1 - (1-\alpha)\gamma}{\alpha\beta_{t-1}}. \tag{28}$$

It is instructive to use the language of macroeconomics to characterize the dynamics. We can interpret $\Delta\mathcal{P}_{t-1}$ as a population *stock* in terms of the population composition and the Markov location strategy in (25) as a population *flow*. Yet, this system is different from standard dynamical systems in that the fixed point of the system in (26) keeps moving. Further, the fixed point depends on F_N, that is, the nature of the population distributions, *i.e.*, what combination (F_B, F_R) can take among the four possible combinations. In particular, the equilibrium existence condition $\widehat{p}_N(N, \emptyset) < 1$ in Proposition 5 does not necessarily imply $1 - F_N > 0$ based on the true distributions; that is, $1 - F_N < 0$ can arise if $(F_B, F_R) = (F_Y, F_Y)$ so that $F_N = F_Y(k) + F_Y(k)$ – which is a necessary condition for $1 - F_N < 0$ (*i.e.*, even for $(F_B, F_R) = (F_Y, F_Y)$, we could have $1 - F_N > 0$). Hence, even when $\Delta\mathcal{P}_{t-1} > 0$, in the next period, $\Delta\mathcal{P}_t < 0$ can arise.[29] Since the fixed point is moving, as it is, we can generalize the no-interior-absorbing state in Section 3 for any initial population difference, and, more importantly, with the moving fixed point, we can relate the size of the exogenous shock with a corresponding initial population difference so that if the shock is small, a large difference in the population size of the two groups is required to show polarization; if the shock is large, a small difference is sufficient.[30]

As shown in the following lemma, the sum of the two probabilities $\mathcal{P}^B_t$ and $\mathcal{P}^R_t$ evolves

[29]If $1 - F_N = 0$ in (26), there is no endogenous part in the dynamics; this is a trivial and uninteresting case.

[30]Note that the fixed point of Section 3 is not moving without exogenous force.

as follows:

$$A_t = \begin{cases} \alpha[1 - F_B(k) + F_R(k)] + (1-\alpha)A_{t-1} & \text{if } \Delta\mathcal{P}_{t-1} \geq 0, \\ \alpha[1 + F_B(k) - F_R(k)] + (1-\alpha)A_{t-1} & \text{if } \Delta\mathcal{P}_{t-1} \leq 0. \end{cases} \tag{29}$$

With each fixed point being dependent on β_{t-1} and the quadratic functional form of A_{t-1} in the denominator of β_{t-1} in (27), Lemma 4 shows that if an initial population difference based on A_0 is not in the middle, a sequence of fixed points satisfies monotonicity; thus z_t^* strictly increases.

Lemma 4 *Suppose (A1)-(A6) and a Markov location equilibrium. A sequence of fixed points satisfies monotonicity such that if* $A_0 < 1 - |F_B(k) - F_R(k)|$ *or* $A_0 > 1 + |F_B(k) - F_R(k)|$, *then* $z_t^* > z_{t-1}^*$ *for all* $t = 1, 2, ...$

Hence, if the two groups have the same type distribution, $F_B = F_R$, then the monotonicity of the fixed point holds for *any* initial value of A_0.

As depicted in Figure 6, whether or not the society converges to a perfectly polarized outcome hinges on the relationship between the fixed point for the real dynamics function in period $t-1$ and the state variable $\Delta\mathcal{P}_{t-1}$ representing the difference in the proportion of group members in location E in that period.[31] By Lemma 4, the fixed point is strictly increasing, so if the state variable in a given period is *lower* than that period's fixed point, the dynamics converge to the completely mixed population composition. Hence, the monotonicity in Lemma 4 provides a consistent pattern for the moving fixed point, but in a way that it is not "helping" the limiting outcome for polarization.

6.1 Completely mixed state

We are now ready to characterize the limiting properties of the dynamics of our system. Due to the shape of this convex dynamical system, we can have either a completely mixed equilibrium or a completely polarized outcome.

We start with the first case and then in the next two subsections, we consider the polarized case.

Proposition 6 *(Completely mixed state) Suppose (A1)-(A6) and a robust location equilibrium. Then, there exists a dynamic spatial equilibrium in Definition 1 for* $t = 1, 2, ...$ *such that any equilibrium location strategy is a Markov strategy in Proposition 5, and further, for* A_0 *satisfying Lemma 4, at any period* $\tau - 1 \geq 0$, *if* $\alpha\beta_{\tau-1}\Delta\mathcal{P}_{\tau-1} + (1-\alpha)\gamma - 1 \leq 0$, *the society converges to a completely mixed state such that* $\lim_{t\geq\tau} \Delta\mathcal{P}_t = 0$.

[31] That is, in the left-panel of Figure 6, $\Delta\mathcal{P}_{t-1}$ is smaller than z_{t-1}^*. Likewise, in the right-panel of the figure, what matters is whether $\Delta\mathcal{P}_{t-1}$ is smaller than z_{t-1}^* or not, but in this oscillation case, the fixed point arises in its negative side. In addition, in the right-panel, note that by $1 - F_N < 0$, $\beta_{t-1} < 0$ in (27).

The most interesting case for the completely mixed state is when there is no systematic polarization, $\gamma = 1$ (or $\epsilon = 0$). That is, if all location choices are made strategically without any exogenous force, then the society converges to a completely mixed state, despite the square functional form implying two different possible convergent results. To show this result, we examine $\beta_{\tau-1}\Delta\mathcal{P}_{\tau-1}$. Using $A_{t-1} \equiv \mathcal{P}^B_{t-1} + \mathcal{P}^R_{t-1}$ in (24) and β_{t-1} in (27), we can write

$$\beta_{\tau-1}\Delta\mathcal{P}_{\tau-1} = \frac{1-F_N}{A_{\tau-1}(2-A_{\tau-1})}\Delta\mathcal{P}_{\tau-1} = (1-F_N)\frac{\mathcal{P}^B_{\tau-1} - \mathcal{P}^R_{\tau-1}}{(\mathcal{P}^B_{\tau-1} + \mathcal{P}^R_{\tau-1})(2-\mathcal{P}^B_{\tau-1} - \mathcal{P}^R_{\tau-1})},$$

where for any $\mathcal{P}^B_{\tau-1}, \mathcal{P}^R_{\tau-1} \in [0,1]$, the fraction in the last line above is always less than or equal to 1. Hence, for each $\tau \geq 1$, we have $\Delta\mathcal{P}_{\tau-1} < \frac{1}{\beta_{\tau-1}} = z^*_{\tau-1}$. Each period's state variable $\Delta\mathcal{P}_{\tau-1}$, representing the difference in the moving probabilities of the two groups, is always smaller than the corresponding fixed point $z^*_{\tau-1}$ of the system. This also entails that z^* in Section 3 is always greater than 1. The intuition is that despite the square function, the difference between two moving probabilities is *generically* lower than the "base" from their sum, $A_{\tau-1}(2-A_{\tau-1})$.

6.2 Polarization: same type distribution

We first present our polarization results for the case of the same type distributions, and in the next subsection, we consider the case of different type distributions. Suppose the two groups have the same type distribution, $F_B = F_R$. Then, from (29), the sum of the two moving probabilities $\mathcal{P}^B_t$ and $\mathcal{P}^R_t$ is given by $A_t = \alpha + (1-\alpha)A_{t-1}$. Without loss of generality, we consider $\Delta\mathcal{P}_0 > 0$ in what follows.[32]

By Proposition 6, to have the divergence result or polarization, the state variable must be greater than the fixed point not only in the initial period but also in *every* subsequent period. In other words, since the fixed point is moving, in particular, increasing toward 1, the state variable must *outgrow* the fixed point; the fixed point can never catch up to the state variable. With systematic polarization, the critical condition is to have a sufficiently large initial difference for the state variable to avoid the "catching-up" possibility as depicted in the left panel of Figure 7. In other words, for a higher degree of systematic polarization γ, we will have a lower sequence of fixed points z^*_t in (28), so there exists a corresponding initial condition $\Delta\mathcal{P}_0$ for which the convergent outcome is complete polarization. The *no-catching up condition* can be found from the sequence of inverse functions of f_{t-1} for all t. We emphasize that $(1-\alpha)\gamma$ in the result includes both a trivial case and a *non-trivial case* $(1-\alpha)\gamma < 1$ – in the sense that the exogenous force itself cannot yield polarization – and what is interesting is, of course, the latter.[33]

[32]If, on the other hand, we have a negative $\Delta\mathcal{P}_0 < 0$, *i.e.* $\mathcal{P}^B_0 < \mathcal{P}^R_0$, then in the first period, there are more B group members in the *West* from $1-\mathcal{P}^B_0 > 1-\mathcal{P}^R_0$, so the same analysis applies, now, in terms of a B player's choice to locate in the *West*.

[33]That is, if $(1-\alpha)\gamma \geq 1$, the exogenous transition *by itself* makes the society converge to the polarization

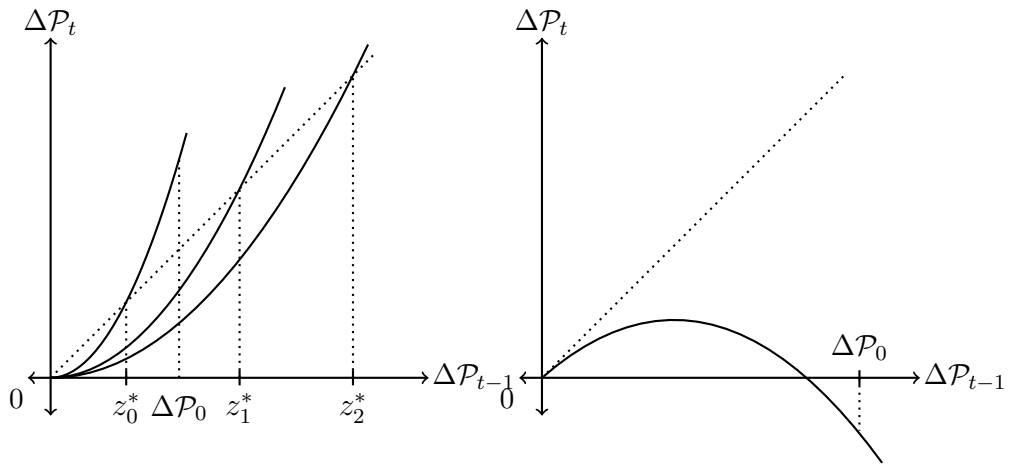

Figure 7: Real dynamics for polarization if $1 - F_N > 0$ (left) and $1 - F_N < 0$ (right)

Proposition 7 *(Polarization: same distribution) Suppose (A1)-(A6) and a robust location equilibrium. Then, there exists a dynamic spatial equilibrium in Definition 1 for $t = 1, 2, ...$ such that any equilibrium location strategy is a Markov strategy in Proposition 5, and further, the polarization results are given as follows.*

(i) If $1-F_N > 0$, then for each $(1-\alpha)\gamma$, there exists $z^\dagger < 1$ such that the society converges to a completely polarized state if an initial population difference $\Delta\mathcal{P}_0 > z^\dagger$.

(ii) If $1-F_N < 0$, then for each $(1-\alpha)\gamma$, there exists $z^{\dagger\dagger} < 1$ such that the society converges to a completely polarized state if an initial population difference $\Delta\mathcal{P}_0 > z^{\dagger\dagger}$.

The dynamical system leads to our first polarization result, which provides conditions under which the society becomes perfectly polarized: Members of the Blue (Red) group locate in the East (West), or the opposite case. A second type of polarization arises from *oscillation* (the right panel of Figure 7) only when the underlying pair of true distributions is given as $(F_B, F_R) = (F_Y, F_Y)$; that is, when both groups turn out to be from the "bad" distribution F_Y.[34]

Note that if the amplifying effect γ is relatively small, by (28), the fixed point for each t becomes larger, resulting in higher critical values $z^\dagger$ and $z^{\dagger\dagger}$ for the complete polarization to arise. Hence, the lower the systematic polarization the less likely the polarization outcome is to arise. An extreme case occurs when $\gamma = 1$, the case without any systematic polarization, which results in convergence to the completely mixed state for any initial difference, as discussed earlier. Hence, we need $\gamma > 1$ to demonstrate the possibility of both the completely mixed and polarized convergence outcomes.

---

outcome. Thus, a non-trivial case is when $(1 - \alpha)\gamma < 1$.

[34]From the right panel of Figure 7, for $\Delta\mathcal{P}_0 > 0$, if $\Delta\mathcal{P}_1 < 1$, then $f_1(z) = -\alpha\beta_1 z^2 + (1 - \alpha)\gamma z$ from (26) and so on.

6.3 Polarization: different type distributions

We now make a comparison between the case in the previous section, where the distributions are the same, $F_B = F_R$, with the case where they are different, $F_B \neq F_R$.

The following proposition reveals that the comparison hinges on the total population size, the sum of the populations of groups B and R who are located in the East, which has a maximum value of 2. If the total population size in E is greater than 1, then polarization is more likely to arise when the B group's distribution is the better one, whereas if the total population size in E is less than 1, then polarization is more likely to arise when the B group's distribution turns out to be the worse one. The second case of Proposition 7, oscillation, only arises with $(F_B, F_R) = (F_Y, F_Y)$ – in which case the comparison with different distributions is not possible – so the following result is written only for $1 - F_N > 0$.

Proposition 8 *(Polarization: different distributions) Suppose (A1)-(A6) and a robust location equilibrium. Then, there exists a dynamic spatial equilibrium in Definition 1 for $t = 1, 2, ...$ such that any equilibrium location strategy is a Markov strategy in Proposition 5, and further, if $1 - F_N > 0$, for each $(1 - \alpha)\gamma$, there exists $\widehat{z}^{\dagger} < 1$ for the different distributions $(F_B, F_R) = (F_X, F_Y)$ such that the society converges to a completely polarized state if an initial population difference $\Delta\mathcal{P}_0 > \widehat{z}^{\dagger}$ with the following properties.*

(i) If $A_0 > 1$, the critical value $\widehat{z}^{\dagger}$ for the different distributions $(F_B, F_R) = (F_X, F_Y)$ is lower (resp. higher) than the critical value $z^{\dagger}$ for the same distribution $(F_B, F_R) = (F_Y, F_Y)$ (resp. $(F_B, F_R) = (F_X, F_X)$).

(ii) If $A_0 < 1$, the critical value $\widehat{z}^{\dagger} < 1$ for the different distributions $(F_B, F_R) = (F_X, F_Y)$ is lower (resp. higher) than the critical value $z^{\dagger}$ for the same distribution $(F_B, F_R) = (F_Y, F_Y)$ (resp. $(F_B, F_R) = (F_X, F_X)$).

If $A_0 = \mathcal{P}_0^B + \mathcal{P}_0^R > 1$, given $\Delta\mathcal{P}_0 > 0$, we have $\mathcal{P}_0^B > \frac{1}{2}$, so with more B players in the East, B group's better distribution $(F_B, F_R) = (F_X, F_Y)$ facilitates polarization more, compared with the case where both groups have the same "bad" distribution case. On the other hand, if $A_0 = \mathcal{P}_0^B + \mathcal{P}_0^R < 1$, given $\Delta\mathcal{P}_0 > 0$, we have $\mathcal{P}_0^R < \frac{1}{2}$, which in turn implies that there are more R players in the West. Likewise, this condition together with the R group having the better distribution $(F_B, F_R) = (F_Y, F_X)$ also facilitates polarization.

7 Discussion

We can expand upon our results in five dimensions. First, a unique equilibrium for the old player in homogeneous matches can be guaranteed for Proposition 4 if (A3) is strengthened such that for each $\theta' > \theta$ in Γ, $d(\theta') - d(\theta) \geq \pi[F_X(\theta') - F_X(\theta)] + (1 - \pi)[F_Y(\theta') - F_Y(\theta)]$ for all $\pi = \pi(I), \pi(N)$. In this case, the relationship holds not only for the unbiased belief but also for $\pi(I), \pi(N)$. This implies that $d(\theta') - d(\theta) \geq a[F_X(\theta') - F_X(\theta)] + a[F_Y(\theta') - F_Y(\theta)]$ for all $\frac{a}{1-a} \leq m$, where interestingly, m is the same as the one from (A6).

Second, while it is reasonable to assume that a young members' beliefs are not inherited from their parent, especially with overlapping generation type dynamics, the $t-1^{th}$ period young might be able to uncover the true distributions if their parents conveyed this information to them and they truly believed what their parents told them.[35] Such an extreme case is when there is oscillation, which can arise only for $(F_B, F_R) = (F_Y, F_Y)$. Except for this particular case, other possibilities that reveal the true distributions can be eliminated by adopting a random proportion of the endogenous location choices such that in each period $\widetilde{\alpha}_{t-1}$ is randomly drawn, where the condition for the Markov strategy is now based on $\alpha = \mathbb{E}[\widetilde{\alpha}_{t-1}]$ and the convergence and polarization results in Propositions 6 & 7 are governed by $f_{t-1}(z, \widetilde{\alpha}_{t-1})$ with a slight abuse of notation, not $f_{t-1}(z)$ in (26).

Third, it is also possible to add population growth, $\delta > 1$ such that for each period $t = 1, 2, ...$, the number of members in each group grows following $L_t = \delta L_{t-1}$ for $L_0 = 1$; that is, the previous unit mass constant population size is now the initial population size. With the population growth factor $\delta > 1$, the number of B or R members moving to E is written as $L_t^B \equiv \mathcal{P}_t^B L_t$ and $L_t^R \equiv \mathcal{P}_t^R L_t$. Then, by denoting $\Delta L_t \equiv L_t^B - L_t^R = \Delta \mathcal{P}_t L_t$, for $\Delta \mathcal{P}_{t-1} \geq 0$, the real population dynamics in (26) can be extended such that $\Delta L_t = \Delta P_t \delta L_{t-1} = \left[\delta \alpha \beta_{t-1} (\Delta \mathcal{P}_{t-1})^2 + \delta(1-\alpha)\gamma \Delta \mathcal{P}_{t-1}\right] L_{t-1}$, which leads to

$$\Delta L_t - \Delta L_{t-1} = \left[\delta \alpha \beta_{t-1} \Delta \mathcal{P}_{t-1} + \delta(1-\alpha)\gamma - 1\right] \Delta L_{t-1}.$$

While the addition of population growth is interesting, the essential part of the dynamics is still controlled by $\Delta \mathcal{P}_{t-1}$. In other words, if $\Delta \mathcal{P}_{t-1}$ converges to zero, then population growth alone cannot reverse this direction.[36]

Fourth, we have not yet discussed what happens *after complete polarization has occurred*. If all B members are in E and all R members are in W, then in each location, with $q_{t-1} = 1$ – without the history $\emptyset$ – the homogeneous match equilibrium in (19) changes to:

$$\begin{aligned} d(k_t^S(\omega^S)) = \pi(\omega^S)[(1 - F_X(k))F_X(k_t^S(I)) + F_X(k)F_X(k_t^S(N))] \\ + (1 - \pi(\omega^S))[(1 - F_Y(k))F_Y(k_t^S(I)) + F_Y(k)F_Y(k_t^S(N))], \end{aligned}$$

which provides the expected payoff for history I and N from meeting a member of the same group by remaining in the same location. On the other hand, moving to the other location yields the expected payoff (2) from meeting with the other group member. Since the comparison between the two mappings in Figure 4 still holds in the limit, that is, $\lim_{t \to \infty} q_{t-1} = 1$, we have $k(I) < k < k(N)$. Since a lower stage game threshold means a higher payoff, as observed previously, those with a *bad* experience from meeting the same group member have

[35] That is, in period $t-1$, their parent could tell them, "When I was young, the difference was $\Delta \mathcal{P}_{t-2}$, and now it is $\Delta \mathcal{P}_{t-1}$, so you see the truth."

[36] To elaborate, this is true for a *non-trivial* case $\delta(1-\alpha)\gamma < 1$ such that the exogenous move by itself cannot generate the polarization; that is, trivially, by having a large δ, a change from $(1-\alpha)\gamma < 1$ with a low initial difference to $\delta(1-\alpha)\gamma \geq 1$ can reverse the direction, without the endogenous part.

an incentive to move to the other location to meet the other group member, given the unbiased belief $\frac{1}{2}$. However, in order to have $\Delta\mathcal{P}_{t-1}$ greater than the fixed point z^*_{t-1} for the polarization in Proposition 7, a *necessary* condition is that $z^*_{t-1} = \frac{1-(1-\alpha)\gamma}{\alpha\beta_{t-1}} < 1$, which can be rewritten as $\frac{1-(1-\alpha)\gamma}{\alpha(1-F_N)} < A_{t-1}(2-A_{t-1}) \leq 1$. This in turn implies that for $\Delta\mathcal{P}_{t-1} = 1$ and $A_{t-1} = 1$, $\alpha(1-F_N)+(1-\alpha)\gamma > 1$ in (26), so the minimum of $\alpha(1-F_N)+(1-\alpha)\gamma$ and 1 is still 1: The endogenous move $\alpha(1-F_N)$ combined with the exogenous force works to maintain the completely polarized outcome. In other words, despite the endogenous decisions, the amplification means that no young player wishes to explore interaction with members of the other group.

Finally, we can weaken (A6) such that it is satisfied only *along* the equilibrium path. This requires a more detailed explanation about how to relate the two different mappings in Section 5. Despite the challenges with the different dimensions, we employ a simple yet clever method to connect them: We construct an auxiliary mapping by parameterizing $p_I(\omega^g)$ and $p_N(\omega^g)$ in (5) for $\lambda \in [0,1]$ in X^A_t and Y^A_t from heterogeneous matches such that for the part with X^A_t,

$$\begin{aligned} p^X_I(\omega^B,\lambda) &\equiv [1-(1-\pi(\omega^B))\lambda](1-F_X(k)) + (1-\pi(\omega^B))\lambda(1-F_Y(k)), \\ p^X_N(\omega^B,\lambda) &\equiv [1-(1-\pi(\omega^B))\lambda]F_X(k) + (1-\pi(\omega^B))\lambda F_Y(k); \end{aligned} \tag{30}$$

and for the part with Y^A_t,

$$\begin{aligned} p^Y_I(\omega^B,\lambda) &\equiv \pi(\omega^B)\lambda(1-F_X(k)) + [1-\pi(\omega^B)\lambda](1-F_Y(k)), \\ p^Y_N(\omega^B,\lambda) &\equiv \pi(\omega^B)\lambda F_X(k) + [1-\pi(\omega^B)\lambda]F_Y(k). \end{aligned} \tag{31}$$

Then, as one can find in the proof of Proposition 4, if $\lambda = 0$, we have the symmetric model, whereas if $\lambda = 1$, we have the asymmetric model, and we aim to make the auxiliary mapping increase in λ. As such, the monotone comparative statics idea of this paper is closely related to Milgrom and Shannon (1994) through Tarski (1955), but it differs from their paper in that the parameter λ is not from the model but is *devised* to connect two functions in the spirit of Homotopy. The auxiliary mapping's equilibrium profile of the old player's thresholds is $(\widehat{k}^\lambda_t(\omega^A))_{\{\omega^A\in\Omega^A\}}$, and a profile including the youthful k threshold is written as $\widehat{\mathbf{k}}^\lambda_t \equiv \big(k, (\widehat{k}^\lambda_t(\omega^A))_{\{\omega^A\in\Omega^A\}}\big)$. Then, a weaker version of (A6) showing the monotonicity along the equilibrium path can be suggested as

(A6′) For each $\lambda > 0$, $F_X(\widehat{k}^\lambda_t(N|\emptyset))) - F_X(\widehat{k}^\lambda_t(I|\emptyset)) \geq m\big[F_Y(\widehat{k}^\lambda_t(N|\emptyset)) - F_Y(\widehat{k}^\lambda_t(I|\emptyset))\big]$.

The condition is satisfied for a large class of parameters. For instance, initially for the homogeneous match, suppose $F_X(k^S_t(N)) - F_X(k^S_t(I)) \geq m\big[F_Y(k^S_t(N)) - F_Y(k^S_t(I))\big]$. If F_X is convex but F_Y is concave on the effective support $[k_X, k_Y]$, then (A6′) holds.[37]

[37] For each λ, $\frac{F_X(\widehat{k}^\lambda_t(N|\emptyset)))-F_X(\widehat{k}^\lambda_t(I|\emptyset))}{\widehat{k}^\lambda_t(N|\emptyset))-\widehat{k}^\lambda_t(I|\emptyset))} \geq \frac{F_X(k^S_t(N))-F_X(k^S_t(I))}{k^S_t(N)-k^S_t(I)}$ and $\frac{F_Y(k^S_t(N))-F_Y(k^S_t(I))}{k^S_t(N)-k^S_t(I)} \geq \frac{F_Y(\widehat{k}^\lambda_t(N|\emptyset)))-F_Y(\widehat{k}^\lambda_t(I|\emptyset)))}{\widehat{k}^\lambda_t(N|\emptyset))-\widehat{k}^\lambda_t(I|\emptyset))}$.

8 Concluding Remarks

What is the origin for the polarization that we often observe in societies by race, language, politics, religion, or other factors? Perhaps the simplest explanation is a preference-based theory, wherein players have tastes for interacting with others who are similar to themselves *e.g.*, as it yields them higher utility and/or lower costs. An alternative but related view allows for some type of special communication or coordination facility with members of one's own group. In this paper we have provided a new and different "statistical discrimination" explanation for understanding the sorting of groups to different locations. Our environment involves players belonging to one of two groups, Red or Blue, where group membership is publicly identifiable. Importantly, there is uncertainty over the distribution of player types for each group, Red and Blue, and private monitoring. Groups are ex-ante equally distributed in terms of ability. Agents live two periods and initially live in one of two locations, the one in which they were born. They interact only with members of their own generation in play of an investment stage game. Based on the histories of play of that game when young, they decide where they will play the game again when they are old. This location choice determines the relative probabilities of meeting other agents from either group in old age, and those matching probabilities serve as the long-term memory of the system.

We assume that agents are not born with any biases favoring their own group or disfavoring the other group. Further, they possess no special facilities for communicating or coordinating with other group members but do inherit the location for play of the investment game from their parents. Agents are initially dispersed between the two locations, and there is a systematic, exogenous, and amplifying force impacting on location decisions, which we attribute to social media or other external influences that are independent of group identity. Starting from these conditions, we show that under certain initial conditions and assuming rational belief updating, the long-term equilibrium outcome of this setup can be that the population becomes perfectly polarized, with all Red and Blue group members choosing separate and distinct locations, and this is a sustained equilibrium outcome.

We emphasize that this outcome is obtained even if the two groups have the *same* type distribution. The equilibrium construction requires that players form perceptions of others' perceptions in deciding where to locate and whether to invest. Still, we show that convergence to the completely polarized outcome is not preordained, but depends crucially on the size of an exogenous force; if this amplifying force is sufficiently small, then the long-term outcome of the system is more likely to be a completely mixed state (no polarization). This finding suggests that policy interventions aimed at reducing the amplifying effects of polarizing forces, *e.g.* social media, may be effective in making the polarization outcome less likely. We leave such analyses to future research.

Appendix A Proofs

Proof of Lemma 1. Each B player chooses location E with probability $\mathcal{P}_t^B$ in period t, so, in the E, the probability that a B player is matched with a B player in period t is

$$\mathcal{P}_t^B \left(\frac{\mathcal{P}_t^B}{\mathcal{P}_t^B + \mathcal{P}_t^R} \right),$$

and each B player chooses location W with probability $1 - \mathcal{P}_t^B$ in period t, so in the W, the probability that a B player is matched with a B player in period t is

$$(1 - \mathcal{P}_t^B) \left(\frac{1 - \mathcal{P}_t^B}{2 - \mathcal{P}_t^B - \mathcal{P}_t^R} \right).$$

Hence, the overall probability that a B player is matched with a B player in period t in (4) is

$$q_t = \frac{(\mathcal{P}_t^B)^2}{\mathcal{P}_t^B + \mathcal{P}_t^R} + \frac{(1 - \mathcal{P}_t^B)^2}{2 - \mathcal{P}_t^B - \mathcal{P}_t^R} = \frac{\mathcal{P}_t^B + \mathcal{P}_t^R - 2\mathcal{P}_t^B \mathcal{P}_t^R}{(\mathcal{P}_t^B + \mathcal{P}_t^R)(2 - \mathcal{P}_t^B - \mathcal{P}_t^R)}.$$

The same formula is also the overall probability that an R player is matched with an R player in period t. On the other hand, the overall probability that a B player is matched with an R player in period t is

$$1 - q_t = \frac{\mathcal{P}_t^B + \mathcal{P}_t^R - (\mathcal{P}_t^B)^2 - (\mathcal{P}_t^R)^2}{(\mathcal{P}_t^B + \mathcal{P}_t^R)(2 - \mathcal{P}_t^B - \mathcal{P}_t^R)}.$$

The same formula is also the overall probability that an R player is matched with a B player in period t. Taking the difference of these two probabilities, we find that

$$q_t - (1 - q_t) = \frac{\left(\mathcal{P}_t^B - \mathcal{P}_t^R\right)^2}{(\mathcal{P}_t^B + \mathcal{P}_t^R)(2 - \mathcal{P}_t^B - \mathcal{P}_t^R)} > 0,$$

so long as $\mathcal{P}_t^B \neq \mathcal{P}_t^R$. ■

Proof of Proposition 1. *Part 1.* First, alternatively, we derive (12) using the general $P_t^B(\omega, \ell_t^B)$ and $P_t^R(\omega, \ell_t^R)$. With (5), in general, $P_t^B(\omega, \ell_t^B)$ and $P_t^R(\omega, \ell_t^R)$ can be derived such that

$$\begin{aligned} \mathbb{E}\left[\mathbf{1}_{\{\ell_t^B(\widetilde{\omega})=E\}} \mid \omega\right] = {} & q_{t-1} p_I(\omega^B) \mathbf{1}_{\{\ell_t^B(I,\emptyset)=E\}} + (1 - q_{t-1}) p_N(\omega^R) \mathbf{1}_{\{\ell_t^B(\emptyset,N)=E\}} \\ & + q_{t-1} p_N(\omega^B) \mathbf{1}_{\{\ell_t^B(N,\emptyset)=E\}} + (1 - q_{t-1}) p_I(\omega^R) \mathbf{1}_{\{\ell_t^B(\emptyset,I)=E\}}, \end{aligned} \tag{32}$$

and

$$\begin{aligned} \mathbb{E}\left[\mathbf{1}_{\{\ell_t^R(\widetilde{\omega})=E\}} \mid \omega\right] = {} & (1 - q_{t-1}) p_I(\omega^B) \mathbf{1}_{\{\ell_t^R(I,\emptyset)=E\}} + q_{t-1} p_N(\omega^R) \mathbf{1}_{\{\ell_t^R(\emptyset,N)=E\}} \\ & + (1 - q_{t-1}) p_N(\omega^B) \mathbf{1}_{\{\ell_t^R(N,\emptyset)=E\}} + q_{t-1} p_I(\omega^R) \mathbf{1}_{\{\ell_t^R(\emptyset,I)=E\}}. \end{aligned} \tag{33}$$

As explained about how to construct $\mathbb{E}\left[\mathbf{1}_{\{\ell_t^B(\widetilde{\omega})=E\}} \mid \omega\right]$ in (32) in Subsection 2.3, one can construct $\mathbb{E}\left[\mathbf{1}_{\{\ell_t^R(\widetilde{\omega})=E\}} \mid \omega\right]$ in (33). For instance, $(1 - q_{t-1}) p_N(\omega^B) \mathbf{1}_{\{\ell_t^R(N,\emptyset)=E\}}$ means that a

player with a history ω reasons, based on his own experience ω, that with probability $1-q_{t-1}$, an arbitrary R player met a different group member, a B member, in youth; with probability $p_N(\omega^B)$, he observed N from the B group member; and if an R player with $\widetilde{\omega} = (N, \emptyset)$ moves to E, *i.e.*, $\mathbf{1}_{\{\ell_t^R(N,\emptyset)=E\}} = 1$, the portion must be counted for $P_t^R(\omega, \ell_t^R)$.

Now, if we incorporate the binary splitting location strategy into (32) and (33), we obtain (12) and (14). First, it is straightforward to find that $1 - \widehat{p}_N(\omega) > 0$ for all $\omega \in \Omega$ if and only if $1 - \widehat{p}_N(N, \emptyset) > 0$ in Proposition 1. From FOSD in (A2), $p_N(\omega^B)$ is strictly decreasing in $\pi(\omega^g)$, so for $(\omega^B, \omega^R) = (N, \emptyset)$ or $(\omega^B, \omega^R) = (\emptyset, N)$ and for $(\widetilde{\omega}^B, \widetilde{\omega}^R) = (I, \emptyset)$ or $(\widetilde{\omega}^B, \widetilde{\omega}^R) = (\emptyset, I)$, we have

$$\widehat{p}_N(\omega^B, \omega^R) = p_N(\omega^B) + p_N(\omega^R) > p_N(\widetilde{\omega}^B) + p_N(\widetilde{\omega}^R) = \widehat{p}_N(\widetilde{\omega}^B, \widetilde{\omega}^R),$$

where recall $p_N(\omega^g)$ in (5) and $\widehat{p}_N(\omega^B, \omega^R)$ in (13). Then, given $(2q_{t-1} - 1) > 0$ from Lemma 1 and the formula in (14), if $1 - \widehat{p}_N(N, \emptyset) > 0$, there exists a binary splitting location equilibrium in Table 2.

Part 2. We only show (i) $|\Delta\mathcal{P}_t| > |\Delta\mathcal{P}_{t-1}| \Rightarrow |\Delta\mathcal{P}_{t+1}| > |\Delta\mathcal{P}_t|$ since the proof for (ii) is identical.

As in the main text, without loss of generality, we suppose that players believe $\Delta P_t(\omega, \ell_t) > 0$. Even with this, we can still have two different possibilities for the next-period $t+1$ beliefs such that $\Delta P_{t+1}(\omega, \ell_{t+1}) > 0$ and $\Delta P_{t+1(}\omega, \ell_{t+1}) < 0$. Under the former case of beliefs, like (16), we have

$$\Delta\mathcal{P}_{t+1} = \beta\Delta\mathcal{P}_t^2, \tag{34}$$

whereas under the latter case of beliefs,

$$\Delta\mathcal{P}_{t+1} = -\beta\Delta\mathcal{P}_t^2. \tag{35}$$

Since $\Delta\mathcal{P}_t > 0$ from $\Delta P_t(\omega, \ell_t) > 0$,

$$|\Delta\mathcal{P}_{t+1}| - |\Delta\mathcal{P}_t| = |\Delta\mathcal{P}_{t+1}| - \Delta\mathcal{P}_t = \begin{cases} \Delta\mathcal{P}_{t+1} - \Delta\mathcal{P}_t & \text{if } \Delta P_{t+1}(\omega, \ell_{t+1}) > 0, \\ -\Delta\mathcal{P}_{t+1} - \Delta\mathcal{P}_t & \text{if } \Delta P_{t+1}(\omega, \ell_{t+1}) < 0. \end{cases}$$

By incorporating (34) and (35), both result in

$$\begin{aligned} |\Delta\mathcal{P}_{t+1}| - |\Delta\mathcal{P}_t| &= \beta\Delta\mathcal{P}_t^2 - \Delta\mathcal{P}_t \\ &= \big[\beta\Delta\mathcal{P}_t - 1\big]\Delta\mathcal{P}_t. \end{aligned}$$

Since $\Delta\mathcal{P}_t > 0$ again from $\Delta P_t(\omega, \ell_t) > 0$,

$$|\Delta\mathcal{P}_t| - |\Delta\mathcal{P}_{t-1}| = \Delta\mathcal{P}_t - |\Delta\mathcal{P}_{t-1}|.$$

Now, we divide the remaining proof into two cases regarding the sign of $\Delta\mathcal{P}_{t-1}$.

Case 1. $\Delta\mathcal{P}_{t-1} > 0$. Then, $|\Delta\mathcal{P}_t| > |\Delta\mathcal{P}_{t-1}| \Leftrightarrow \Delta\mathcal{P}_t > \Delta\mathcal{P}_{t-1}$, so we have the strict inequality below:

$$\begin{aligned}
|\Delta\mathcal{P}_{t+1}| - |\Delta\mathcal{P}_t| &= \beta\Delta\mathcal{P}_t^2 - \Delta\mathcal{P}_t \\
&= \big[\beta\Delta\mathcal{P}_t - 1\big]\Delta\mathcal{P}_t \\
&> \big[\beta\Delta\mathcal{P}_{t-1} - 1\big]\Delta\mathcal{P}_t \\
&= \left[\frac{\Delta\mathcal{P}_t - \Delta\mathcal{P}_{t-1}}{\Delta\mathcal{P}_{t-1}}\right]\Delta\mathcal{P}_t,
\end{aligned}$$

where the last equality follows from

$$|\Delta\mathcal{P}_t| - |\Delta\mathcal{P}_{t-1}| = \Delta\mathcal{P}_t - \Delta\mathcal{P}_{t-1} = \beta\Delta\mathcal{P}_{t-1}^2 - \Delta\mathcal{P}_{t-1}.$$

Hence,

$$\frac{|\Delta\mathcal{P}_{t+1}| - |\Delta\mathcal{P}_t|}{|\Delta\mathcal{P}_t|} > \frac{|\Delta\mathcal{P}_t| - |\Delta\mathcal{P}_{t-1}|}{|\Delta\mathcal{P}_{t-1}|}.$$

Case 2. $\Delta\mathcal{P}_{t-1} < 0$. Then, $|\Delta\mathcal{P}_t| > |\Delta\mathcal{P}_{t-1}| \Leftrightarrow \Delta\mathcal{P}_t > -\Delta\mathcal{P}_{t-1}$, so we have the strict inequality below:

$$\begin{aligned}
|\Delta\mathcal{P}_{t+1}| - |\Delta\mathcal{P}_t| &= \beta\Delta\mathcal{P}_t^2 - \Delta\mathcal{P}_t \\
&= \big[\beta\Delta\mathcal{P}_t - 1\big]\Delta\mathcal{P}_t \\
&> \big[-\beta\Delta\mathcal{P}_{t-1} - 1\big]\Delta\mathcal{P}_t \\
&= -\left[\frac{\Delta\mathcal{P}_t + \Delta\mathcal{P}_{t-1}}{\Delta\mathcal{P}_{t-1}}\right]\Delta\mathcal{P}_t,
\end{aligned}$$

where the last equality follows from

$$|\Delta\mathcal{P}_t| - |\Delta\mathcal{P}_{t-1}| = \Delta\mathcal{P}_t + \Delta\mathcal{P}_{t-1} = \beta\Delta\mathcal{P}_{t-1}^2 + \Delta\mathcal{P}_{t-1}$$

Hence, given $\Delta\mathcal{P}_{t-1} < 0$,

$$\frac{|\Delta\mathcal{P}_{t+1}| - |\Delta\mathcal{P}_t|}{|\Delta\mathcal{P}_t|} > \frac{|\Delta\mathcal{P}_t| - |\Delta\mathcal{P}_{t-1}|}{|\Delta\mathcal{P}_{t-1}|}.$$

The result is established with the cases above. ■

Proof of Proposition 2. We start by deriving the probability of a matched partner choosing NoInvest $\Pr(N|\omega^S, \mathbf{k}_t^S)$ in (18) precisely, where recall that $\mathbf{k}_t^S \equiv \big(k, (k_t^S(\omega^S))_{\{\omega^S \in \Omega^S\}}\big)$. The probability is given as

$$\begin{aligned}
\Pr(N|\omega^S, \mathbf{k}_t^S) = \Pr(N|\omega^S, \mathbf{k}_t^S) &= \Pr(N, X^S|\omega^S, \mathbf{k}_t^S) + \Pr(N, Y^S|\omega^S, \mathbf{k}_t^S) \\
&= \Pr(X^S|\omega^S, \mathbf{k}_t^S)\Pr(N|X^S, \omega^S, \mathbf{k}_t^S) + \Pr(Y^S|\omega^S, \mathbf{k}_t^S)\Pr(N|Y^S, \omega^S, \mathbf{k}_t^S).
\end{aligned} \tag{36}$$

For each $r \in \{X, Y\}$, $\Pr(r^S|\omega^S, \mathbf{k}_t^S)$ is the probability that the same group's distribution is F_r and $\Pr(N|r^S, \omega^S, \mathbf{k}_t^S)$ is the probability that the partner chooses NoInvest, conditional

on F_r. In particular, the latter probability depends on previous observations of the partner, *i.e.*, ω^S. We denote $X_t^S(\mathbf{k}_t^S) \equiv \Pr(N|X^S, \omega^S, \mathbf{k}_t^S)$ and $Y_t^S(\mathbf{k}_t^S) \equiv \Pr(N|Y^S, \omega^S, \mathbf{k}_t^S)$ and derive them as (17). Since for $\omega^S = I, N \in \Omega^S$, $\Pr(X^S|\omega^S, \mathbf{k}_t^S) = \pi(\omega^S)$ in (3), and no previous match with B player yields no updating on B group, $\pi(\emptyset) = \frac{1}{2}$. Together, we have

$$\Pr(N|\omega^S, \mathbf{k}_t^S) = \pi(\omega^S) X_t^S(\mathbf{k}_t^S) + (1 - \pi(\omega^S)) Y_t^S(\mathbf{k}_t^S).$$

We denote $\Phi_t^S(k_t^S, \omega^S) \equiv d^{-1}\big(\pi(\omega^S) X_t^S(\mathbf{k}_t^S) + (1 - \pi(\omega^S)) Y_t^S(\mathbf{k}_t^S)\big)$. Then, a homogeneous match equilibrium is a fixed point of a mapping $\mathbf{\Phi}_t^S : [\underline{\theta}, \overline{\theta}]^3 \to [\underline{\theta}, \overline{\theta}]^3$ that is defined as

$$\mathbf{\Phi}_t^S(k_t^S) \equiv \Big(\Phi_t^S(k_t^S, \omega^S)\Big)_{\{\omega^S \in \Omega^S\}}. \tag{37}$$

For the characterization, it suffices to show that for any pair of two histories $\widehat{\omega}^S, \omega^S \in \Omega^S$ with $\pi(\widehat{\omega}^S) > \pi(\omega^S)$, we have $k_t(\widehat{\omega}^S) < k_t(\omega^S)$. Suppose, on the contrary, that $k_t(\widehat{\omega}^S) \geq k_t(\omega^S)$. From (19), any pair of two histories $\widehat{\omega}^S, \omega^S \in \Omega^S$ yields a difference in two thresholds such that

$$d(k_t^S(\widehat{\omega}^S)) - d(k_t^S(\omega^S)) = \big(\pi(\widehat{\omega}^S) - \pi(\omega^S)\big)\big[X_t^S(\mathbf{k}_t^S) - Y_t^S(\mathbf{k}_t^S)\big]. \tag{38}$$

We divide the proof into two cases.

Case 1. $k_t^S(\widehat{\omega}^S) = k_t^S(\omega^S)$. The difference in (38) results in $X_t^S(\mathbf{k}_t^S) = Y_t^S(\mathbf{k}_t^S)$, which in turn implies that $k_t = k_t^S(\omega^S)$ for all ω^S. Then, it follows from (17) that $X_t^S(\mathbf{k}_t^S) - Y_t^S(\mathbf{k}_t^S) = F_X(k_t) - F_Y(k_t)$, so

$$\begin{aligned} 0 = d(k_t^S(\widehat{\omega}^S)) - d(k_t^S(\omega^S)) &= \big(\pi(\widehat{\omega}^S) - \pi(\omega^S)\big)\big[X_t^S(\mathbf{k}_t^S) - Y_t^S(\mathbf{k}_t^S)\big] \\ &= \big(\pi(\widehat{\omega}^S) - \pi(\omega^S)\big)[F_X(k_t) - F_Y(k_t)] < 0, \end{aligned}$$

which is a contradiction.

Case 2. $k_t^S(\widehat{\omega}^S) > k_t^S(\omega^S)$. The formula in (38) and $\pi(\widehat{\omega}^S) > \pi(\omega^S)$ lead to $X_t^S(\mathbf{k}_t^S) > Y_t^S(\mathbf{k}_t^S)$. Consider $k_t^{\max} \equiv \max\{k_t^S(I), k_t^S(\emptyset), k_t^S(N)\}$ and $k_t^{\min} \equiv \min\{k_t^S(I), k_t^S(\emptyset), k_t^S(N)\}$. Then, the difference in their thresholds is

$$d\big(k_t^{\max}\big) - d\big(k_t^{\max}\big) = \big(\pi^{\max} - \pi^{\min}\big)\big[X_t^S(\mathbf{k}_t^S) - Y_t^S(\mathbf{k}_t^S)\big].$$

From $X_t^S(\mathbf{k}_t^S) > Y_t^S(\mathbf{k}_t^S)$ and $0 < \pi^{\max} - \pi^{\min} < 1$, for each $r \in \{X, Y\}$, we have

$$d\big(k_t^{\max}\big) - d\big(k_t^{\max}\big) = \big(\pi^{\max} - \pi^{\min}\big)\big[X_t^S(\mathbf{k}_t^S) - Y_t^S(\mathbf{k}_t^S)\big] < \big[X_t^S(\mathbf{k}_t^S) - Y_t^S(\mathbf{k}_t^S)\big] < F_r(k_t^{\max}) - F_r(k_t^{\min}),$$

where the last inequality follows from $k_t^{\max}$ and $k_t^{\min}$. Then, this yields a contradiction with (A3) since (A3) implies that for any pair $\theta' > \theta$ in Γ, $d(\theta') - d(\theta) \geq F_r(\theta') - F_r(\theta)$ for at least one $r \in \{X, Y\}$. ■

Proof of Proposition 3. We start by deriving the probability of a matched partner choosing NoInvest $\Pr(N|\omega^A, \mathbf{k}_t^A)$ in (20) precisely, where recall that $\mathbf{k}_t^A \equiv \big(k, (k_t^A(\omega^A))_{\{\omega^A \in \Omega^A\}}\big)$.

Without loss of generality, let us examine the problem from the B member's perspective to investigate the NoInvest probability $\Pr(N|\omega^A, \mathbf{k}_t^A)$, which is given as

$$\begin{aligned}\Pr(N|\omega^A, \mathbf{k}_t^A) = &\Pr(N, (X^B, X^R)|\omega^A, \mathbf{k}_t^A) + \Pr(N, (X^B, Y^R)|\omega^A, \mathbf{k}_t^A) \\ &+ \Pr(N, (Y^B, X^R)|\omega^A, \mathbf{k}_t^A) + \Pr(N, (Y^B, Y^R)|\omega^A, \mathbf{k}_t^A) \\ = &\Pr(X^R|\omega^A, \mathbf{k}_t^A)\left[\Pr(N, X^B|X^R, \omega^A, \mathbf{k}_t^A) + \Pr(N, Y^B|X^R, \omega^A, \mathbf{k}_t^A)\right] \\ &+ \Pr(Y^R|\omega^A, \mathbf{k}_t^A)\left[\Pr(N, X^B|Y^R, \omega^A, \mathbf{k}_t^A) + \Pr(N, Y^B|Y^R, \omega^A, \mathbf{k}_t^A)\right].\end{aligned} \tag{39}$$

Furthermore, denote

$$\begin{aligned}X_t^A(\mathbf{k}_t^A, \omega^B) &\equiv \Pr(N, X^B|X^R, \omega^A, \mathbf{k}_t^A) + \Pr(N, Y^B|X^R, \omega^A, \mathbf{k}_t^A), \\ Y_t^A(\mathbf{k}_t^A, \omega^B) &\equiv \Pr(N, X^B|Y^R, \omega^A, \mathbf{k}_t^A) + \Pr(N, Y^B|Y^R, \omega^A, \mathbf{k}_t^A).\end{aligned}$$

We now derive $X_t^A(\mathbf{k}_t^A, \omega^B)$ and $Y_t^A(\mathbf{k}_t^A, \omega^B)$ with $p_I(\omega^B)$ and $p_N(\omega^B)$ such that

$$\begin{aligned}X_t^A(\mathbf{k}_t^A, \omega^B) = &\, q_{t-1}(1 - F_X(k))F_X(k_t^A(\emptyset|I)) + q_{t-1}F_X(k)F_X(k_t^A(\emptyset|N)) \\ &+ (1 - q_{t-1})p_I(\omega^B)F_X(k_t^A(I|\emptyset)) + (1 - q_{t-1})p_N(\omega^B)F_X(k_t^A(N|\emptyset)), \\ Y_t^A(\mathbf{k}_t^A, \omega^B) = &\, q_{t-1}(1 - F_Y(k))F_Y(k_t^A(\emptyset|I)) + q_{t-1}F_Y(k)F_Y(k_t^A(\emptyset|N)) \\ &+ (1 - q_{t-1})p_I(\omega^B)F_Y(k_t^A(I|\emptyset)) + (1 - q_{t-1})p_N(\omega^B)F_Y(k_t^A(N|\emptyset)).\end{aligned} \tag{40}$$

For instance, conditional on the R group distribution, X^R or Y^R, with probability q_{t-1} from (4), the matched R player met another R member previously when young, and with probability $(1 - F_r(k))$, he observed Invest from that partner, which with a corresponding threshold $k_t^A(\emptyset|I)$ yields the first term $q_{t-1}(1 - F_X(k))F_X(k_t^A(\emptyset|I))$. In addition, with probability $1 - q_{t-1}$, the matched R player met a B player previously when young, and with probability $p_I(\omega^B)$, he observed Invest from that partner, which with a corresponding threshold $k_t^A(I|\emptyset)$ yields the third term $(1 - q_{t-1})p_I(\omega^B)F_X(k_t^A(I|\emptyset))$. The other two terms can be readily derived accordingly. Since for $\omega^R = I, N$ from $\omega^A \in \Omega^A$, $\Pr(X^R|\omega^R, \mathbf{k}_t^S) = \pi(\omega^R)$ in (3) together with $\pi(\emptyset) = \frac{1}{2}$, we have

$$\Pr(N|\omega^A, \mathbf{k}_t^A) = \pi(\omega^R)X_t^A(\mathbf{k}_t^A, \omega^B) + (1 - \pi(\omega^R))Y_t^A(\mathbf{k}_t^A, \omega^B).$$

Then, extending a homogeneous match equilibrium in (37), a heterogeneous match equilibrium is a fixed point of a mapping $\mathbf{\Phi}_t^A : [\underline{\theta}, \overline{\theta}]^4 \to [\underline{\theta}, \overline{\theta}]^4$ that is defined as

$$\mathbf{\Phi}_t^A(k_t^A) \equiv \left(\Phi_t^A(k_t^A, \omega^A)\right)_{\{\omega^A \in \Omega^A\}}, \tag{41}$$

where $\Phi_t^A(k_t^A, \omega^A) \equiv d^{-1}\big(\pi(\omega^R)X_t^A(\mathbf{k}_t^A, \omega^B) + (1 - \pi(\omega^R))Y_t^A(\mathbf{k}_t^A, \omega^B)\big)$. The lemma below is the first step in this direction.

Lemma 5 *Suppose (A1)-(A4). Then a heterogeneous match equilibrium profile of the old player's thresholds satisfies the following properties: for each $r \in \{X, Y\}$,*

(i) $d(k_t^A(I|\emptyset)) - d(k_t^A(\emptyset|I)) < q_{t-1}(\pi(I^R) - \pi(\emptyset^R))(F_Y(k) - F_X(k))\left[F_r(k_t^A(\emptyset|I)) - F_r(k_t^A(\emptyset|N))\right]$,

(ii) $d(k_t^A(N|\emptyset))-d(k_t^A(\emptyset|N)) > q_{t-1}(\pi(N^R)-\pi(\emptyset^R))(F_Y(k)-F_X(k))\left[F_r(k_t^A(\emptyset|I)) - F_r(k_t^A(\emptyset|N))\right]$,

(iii) $d(k_t^A(\phi|I)) - d(k_t^A(\phi|N))$
$= \frac{(1-q_{t-1})}{2}(\pi(I^B) - \pi(N^B))[F_Y(k) - F_X(k)] \left[\begin{array}{l} F_X(k_t^A(I|\phi)) - F_X(k_t^A(N|\phi)) \\ +F_Y(k_t^A(I|\phi)) - F_Y(k_t^A(N|\phi)) \end{array} \right]$.

To establish a monotonicity result for a heterogeneous match, we first prove that $k_t^A(N|\emptyset) > k_t^A(I|\emptyset)$ in this proof (the proof of Lemma 5 is relegated to Appendix B.3). Once that is shown, then, Lemma 5 (iii) implies that $k_t^A(\emptyset|N) > k_t^A(\emptyset|I)$, which in turn can be incorporated into Lemma 5 (i) and (ii) to obtain $k_t^A(I|\emptyset) < k_t^A(\emptyset|I)$ and $k_t^A(N|\emptyset) > k_t^A(\emptyset|N)$, respectively.

Now, consider the difference between $k_t^A(I|\emptyset)$ and $k_t^A(N|\emptyset)$ such that

$$d(k_t^A(I|\emptyset)) - d(k_t^A(N|\emptyset)) = \big(\pi(I^R) - \pi(N^R)\big)[X_t^A(\mathbf{k}_t^A, \emptyset^B) - Y_t^A(\mathbf{k}_t^A, \emptyset^B)].$$

We first show $k_t^A(N|\emptyset) > k_t^A(I|\emptyset)$. Suppose, on the contrary, that $k_t^A(I|\emptyset) \geq k_t^A(N|\emptyset)$. Then, Lemma 5 (iii) implies that $k_t^A(\emptyset|I) \geq k_t^A(\emptyset|N)$. Then given $k_t^A(I|\emptyset) \geq k_t^A(N|\emptyset)$ and $k_t^A(\emptyset|I) \geq k_t^A(\emptyset|N)$, for $X_t^A(\mathbf{k}_t^A, \emptyset^B)$, by taking those two higher values, and for $Y_t^A(\mathbf{k}_t^A, \emptyset^B)$, by taking those two lower values, $X_t^A(\mathbf{k}_t^A, \emptyset^B) - Y_t^A(\mathbf{k}_t^A, \emptyset^B)$ is rewritten as

$$\begin{aligned}
&\left[\begin{array}{l} q_{t-1}(1 - F_X(k))F_X(k_t^A(\emptyset|I)) + q_{t-1}F_X(k)F_X(k_t^A(\emptyset|N)) \\ +(1 - q_{t-1})\big[\pi(\emptyset^B)(1 - F_X(k)) + (1 - \pi(\emptyset^B))(1 - F_Y(k))\big]F_X(k_t^A(I|\emptyset)) \\ +(1 - q_{t-1})\big[\pi(\emptyset^B)F_X(k) + (1 - \pi(\emptyset^B))F_Y(k)\big]F_X(k_t^A(N|\emptyset)) \end{array} \right] \\
&- \left[\begin{array}{l} q_{t-1}(1 - F_Y(k))F_Y(k_t^A(\emptyset|I)) + q_{t-1}F_Y(k)F_Y(k_t^A(\emptyset|N)) \\ +(1 - q_{t-1})\big[\pi(\emptyset^B)(1 - F_X(k)) + (1 - \pi(\emptyset^B))(1 - F_Y(k))\big]F_Y(k_t^A(I|\emptyset)) \\ +(1 - q_{t-1})\big[\pi(\emptyset^B)F_X(k) + (1 - \pi(\emptyset^B))F_Y(k)\big]F_Y(k_t^A(N|\emptyset)) \end{array} \right] \\
&\leq q_{t-1}F_X(k_t^A(\emptyset|I)) + (1 - q_{t-1})F_X(k_t^A(I|\emptyset)) - [q_{t-1}F_Y(k_t^A(\emptyset|N)) + (1 - q_{t-1})F_Y(k_t^A(N|\emptyset))] \\
&< q_{t-1}F_r(k_t^A(\emptyset|I)) + (1 - q_{t-1})F_r(k_t^A(I|\emptyset)) - [q_{t-1}F_r(k_t^A(\emptyset|N)) + (1 - q_{t-1})F_r(k_t^A(N|\emptyset))]
\end{aligned}$$

for all $r \in \{X, Y\}$, where the last inequality follows from the FOSD between F_X and F_Y. Hence, with $0 < \pi(I^R) - \pi(N^R) < 1$,

$$\begin{aligned}
&d(k_t^A(I|\emptyset)) - d(k_t^A(N|\emptyset)) \\
&< q_{t-1}F_r(k_t^A(\emptyset|I)) + (1 - q_{t-1})F_r(k_t^A(I|\emptyset)) - [q_{t-1}F_r(k_t^A(\emptyset|N)) + (1 - q_{t-1})F_r(k_t^A(N|\emptyset))].
\end{aligned} \tag{42}$$

First, if $k_t^A(I|\emptyset)) = k_t^A(N|\emptyset)$, Lemma 5 (iii) implies $k_t^A(\emptyset|I) = k_t^A(\emptyset|N)$, so we have a contradiction with the above inequality. Now, suppose $k_t^A(I|\emptyset) > k_t^A(N|\emptyset)$. Then, from Lemma 5 (iii),

$$\begin{aligned}
&d(k_t^A(\emptyset|I)) - d(k_t^A(\emptyset|N)) \\
&= \frac{1}{2}(1 - q_{t-1})\big(\pi(I^B) - \pi(N^B)\big)[F_Y(k) - F_X(k)] \left[\begin{array}{l} F_X(k_t^A(I|\emptyset)) - F_X(k_t^A(N|\emptyset)) \\ +F_Y(k_t^A(I|\emptyset)) - F_Y(k_t^A(N|\emptyset)) \end{array} \right] \\
&< d(k_t^A(I|\emptyset)) - d(k_t^A(N|\emptyset)),
\end{aligned}$$

where the last inequality follows from (A3) and $0 < (1-q_{t-1})\big(\pi(I^B)-\pi(N^B)\big)[F_Y(k)-F_X(k)] < 1$. Together, for each $r \in \{X, Y\}$, we have

$$\begin{aligned}
&d(k_t^A(\emptyset|I)) - d(k_t^A(\emptyset|N)) < d(k_t^A(I|\emptyset)) - d(k_t^A(N|\emptyset)) \\
&< q_{t-1}F_r(k_t^A(\emptyset|I)) + (1-q_{t-1})F_r(k_t^A(I|\emptyset)) - [q_{t-1}F_r(k_t^A(\emptyset|N)) + (1-q_{t-1})F_r(k_t^A(N|\emptyset))] \\
&= q_{t-1}[F_r(k_t^A(\emptyset|I)) - F_r(k_t^A(\emptyset|N))] + (1-q_{t-1})[F_r(k_t^A(I|\emptyset)) - F_r(k_t^A(N|\emptyset))] \\
&\leq \max\{F_r(k_t^A(\emptyset|I)) - F_r(k_t^A(\emptyset|N)), F_r(k_t^A(I|\emptyset)) - F_r(k_t^A(N|\emptyset))\},
\end{aligned}$$

which yields a contradiction with (A3). ■

Proof of Lemma 2. From the homogeneous match in (18), in equilibrium, we have

$$U_t^S(\theta, \omega^S, \mathbf{k}_t^S) \equiv d(\theta) - [\pi(\omega^S)X_t^S(\mathbf{k}_t^S) + (1-\pi(\omega^S))Y_t^S(\mathbf{k}_t^S)] = d(\theta) - d(k_t^S(\omega^S)),$$

and from the heterogeneous match in (20), in equilibrium, we have

$$U_t^A(\theta, \omega^A, \mathbf{k}_t^A) \equiv d(\theta) - [\pi(\omega^R)X_t^A(\mathbf{k}_t^A, \pi(\omega^B)) + (1-\pi(\omega^R))Y_t^A(\mathbf{k}_t^A, \pi(\omega^B))] = d(\theta) - d(k_t^A(\omega^A)).$$

The difference yields the result. Note that

$$\frac{P_t^B(\omega)}{P_t^B(\omega)+P_t^R(\omega)} - \frac{1-P_t^B(\omega)}{2-P_t^B(\omega)-P_t^R(\omega)} = \frac{P_t^B(\omega)-P_t^R(\omega)}{(P_t^B(\omega)+P_t^R(\omega))\,(2-P_t^B(\omega)-P_t^R(\omega))}$$

and

$$\frac{P_t^R(\omega)}{P_t^B(\omega)+P_t^R(\omega)} - \frac{1-P_t^R(\omega)}{2-P_t^B(\omega)-P_t^R(\omega)} = \frac{P_t^R(\omega)-P_t^B(\omega)}{(P_t^B(\omega)+P_t^R(\omega))\,(2-P_t^B(\omega)-P_t^R(\omega))}.$$

Hence, the difference in expected payoffs to a B player is given by

$$\frac{P_t^B(\omega)-P_t^R(\omega)}{(P_t^B(\omega)+P_t^R(\omega))\,(2-P_t^B(\omega)-P_t^R(\omega))}\left[U_t^S(\theta, \omega^S, \mathbf{k}_t^S) - U_t^A(\theta, \omega^A, \mathbf{k}_t^A)\right].$$

The result follows. ■

Proof of Proposition 4. Despite the challenges with respect to comparing thresholds from the two different mappings, $\mathbf{\Phi}_t^S : [\underline{\theta}, \overline{\theta}]^3 \to [\underline{\theta}, \overline{\theta}]^3$ in (37) and $\mathbf{\Phi}_t^A : [\underline{\theta}, \overline{\theta}]^4 \to [\underline{\theta}, \overline{\theta}]^4$ in (41), we connect them by constructing an *auxiliary* mapping $\widehat{\mathbf{\Phi}}_t(\cdot, \lambda) : [\underline{\theta}, \overline{\theta}]^4 \to [\underline{\theta}, \overline{\theta}]^4$ connecting them. Specifically, we parameterize $p_I(\omega^B)$ and $p_N(\omega^B)$ in (5) with $\lambda \in [0, 1]$ such as $p_I^X(\omega^B, \lambda)$ and $p_N^X(\omega^B, \lambda)$ in (30) and $p_I^Y(\omega^B, \lambda)$ and $p_N^Y(\omega^B, \lambda)$ in (31).

Part 1. We first show (i). With (30) and (31), define

$$\begin{aligned}
\widehat{X}_t(\widehat{\mathbf{k}}_t^\lambda, \omega^B, \lambda) \equiv\ & q_{t-1}(1-F_X(k))F_X(\widehat{k}_t^\lambda(\emptyset|I)) + q_{t-1}F_X(k)F_X(\widehat{k}_t^\lambda(\emptyset|N)) \\
&+ (1-q_{t-1})p_I^X(\omega^B, \lambda)F_X(\widehat{k}_t^\lambda(I|\emptyset)) + (1-q_{t-1})p_N^X(\omega^B, \lambda)F_X(\widehat{k}_t^\lambda(N|\emptyset)), \\
\widehat{Y}_t(\widehat{\mathbf{k}}_t^\lambda, \omega^B, \lambda) \equiv\ & q_{t-1}(1-F_Y(k))F_Y(\widehat{k}_t^\lambda(\emptyset|I)) + q_{t-1}F_Y(k)F_Y(\widehat{k}_t^\lambda(\emptyset|N)) \\
&+ (1-q_{t-1})p_I^Y(\omega^B, \lambda)F_Y(\widehat{k}_t^\lambda(I|\emptyset)) + (1-q_{t-1})p_N^Y(\omega^B, \lambda)F_Y(\widehat{k}_t^\lambda(N|\emptyset)).
\end{aligned}$$

Together, the auxiliary mapping is defined as

$$\widehat{\mathbf{\Phi}}_t(\widehat{k}_t^\lambda, \lambda) \equiv \left(\widehat{\Phi}_t(\widehat{k}_t^\lambda, \omega^A, \lambda)\right)_{\{\omega^A \in \Omega^A\}}, \tag{43}$$

where $\widehat{\Phi}_t(\widehat{k}_t^\lambda, \omega^A, \lambda) \equiv d^{-1}\big(\pi(\omega^R)\widehat{X}_t(\widehat{\mathbf{k}}_t^\lambda, \omega^B, \lambda) + (1-\pi(\omega^R))\widehat{Y}_t(\widehat{\mathbf{k}}_t^\lambda, \omega^B, \lambda)\big)$. If $\lambda = 0$, the auxiliary mapping becomes the homogeneous mapping with $\widehat{\Phi}_t(\widehat{k}_t^\lambda, \emptyset|I, \lambda) = \widehat{\Phi}_t(\widehat{k}_t^\lambda, \emptyset|N, \lambda)$, so *effectively*, there are three functions. On the other hand, if $\lambda = 1$, it becomes the heterogeneous mapping. Since $\widehat{\Phi}_t(\widehat{k}_t^\lambda, \omega^A, \lambda)$ is monotone in $\widehat{k}_t^\lambda$, it suffices to show that the mapping is increasing in λ in order to apply the monotone comparative statics in Milgrom and Shannon (1994) to this parametrized approach.

Step 1. Show $\widehat{k}_t^\lambda(N|\emptyset) > \widehat{k}_t^\lambda(I|\emptyset)$ for all $\lambda > 0$. To do that, we prove (iii) of Lemma 5 for the auxiliary mapping in (43) given $\lambda > 0$. We obtain a formula that is similar to the one from the proof for (iii) of Lemma 5 except for λ in the formula below. By taking the difference between $\widehat{k}_t^\lambda(\emptyset|I)$ and $\widehat{k}_t^\lambda(\emptyset|N)$,

$$\begin{aligned}
&d(\widehat{k}_t^\lambda(\emptyset|I)) - d(\widehat{k}_t^\lambda(\emptyset|N)) \\
&= \pi(\emptyset^R)\widehat{X}_t(\widehat{\mathbf{k}}_t^\lambda, I^B, \lambda) + (1-\pi(\emptyset^R))\widehat{Y}_t(\widehat{\mathbf{k}}_t^\lambda, I^B, \lambda) - [\pi(\emptyset^R)\widehat{X}_t(\widehat{\mathbf{k}}_t^\lambda, N^B, \lambda) + (1-\pi(\emptyset^R))\widehat{Y}_t(\widehat{\mathbf{k}}_t^\lambda, I^B, \lambda)] \\
&= \pi(\emptyset^R)(1-q_{t-1})\left[F_X(\widehat{k}_t^\lambda(I|\emptyset)) - F_X(\widehat{k}_t^\lambda(N|\emptyset))\right]\lambda\big(\pi(I^B) - \pi(N^B)\big)[F_Y(k) - F_X(k)] \\
&\quad + (1-\pi(\emptyset^R))(1-q_{t-1})\left[F_Y(\widehat{k}_t^\lambda(I|\emptyset)) - F_Y(\widehat{k}_t^\lambda(N|\emptyset))\right]\lambda\big(\pi(I^B) - \pi(N^B)\big)[F_Y(k) - F_X(k)] \\
&= \frac{1}{2}\lambda(1-q_{t-1})\big(\pi(I^B) - \pi(N^B)\big)[F_Y(k) - F_X(k)]\begin{bmatrix} F_X(\widehat{k}_t^\lambda(I|\emptyset)) - F_X(\widehat{k}_t^\lambda(N|\emptyset)) \\ +F_Y(\widehat{k}_t^\lambda(I|\emptyset)) - F_Y(\widehat{k}_t^\lambda(N|\emptyset)) \end{bmatrix},
\end{aligned}$$

which establishes Lemma (iii) version of the auxiliary mapping. Now, we are ready to show $\widehat{k}_t^\lambda(N|\emptyset) > \widehat{k}_t^\lambda(I|\emptyset)$. Consider the difference $\widehat{k}_t^\lambda(I|\emptyset)$ and $\widehat{k}_t^\lambda(N|\emptyset)$ such that

$$d(\widehat{k}_t^\lambda(I|\emptyset)) - d(\widehat{k}_t^\lambda(N|\emptyset)) = \big(\pi(I^R) - \pi(N^R)\big)[\widehat{X}_t(\widehat{\mathbf{k}}_t^\lambda, \emptyset^B, \lambda) - \widehat{Y}_t(\widehat{\mathbf{k}}_t^\lambda, \emptyset^B, \lambda)].$$

Suppose, on the contrary, that $\widehat{k}_t^\lambda(I|\emptyset) \geq \widehat{k}_t^\lambda(N|\emptyset)$. Then, Lemma (iii) version of the auxiliary mapping implies that $\widehat{k}_t^\lambda(\emptyset|I) \geq \widehat{k}_t^\lambda(\emptyset|N)$. Hence, given $\widehat{k}_t^\lambda(I|\emptyset) \geq \widehat{k}_t^\lambda(N|\emptyset)$ and $\widehat{k}_t^\lambda(\emptyset|I) \geq \widehat{k}_t^\lambda(\emptyset|N)$, for $\widehat{X}_t(\widehat{\mathbf{k}}_t^\lambda, \emptyset^B, \lambda)$, by taking those two higher values, and for $\widehat{Y}_t(\widehat{\mathbf{k}}_t^\lambda, \emptyset^B, \lambda)$, by taking those two lower values, λ disappears, so $\widehat{X}_t(\widehat{\mathbf{k}}_t^\lambda, \emptyset^B, \lambda) - \widehat{Y}_t(\widehat{\mathbf{k}}_t^\lambda, \emptyset^B, \lambda)$ is rewritten as exactly the same as the one in the proof of Proposition 3, which leads to (42). Then, we reach the same contradiction.

Step 2. Then for each $\omega^A \in \Omega^A$, the derivative of $\widehat{\Phi}_t(\widehat{k}_t^\lambda, \omega^A, \lambda)$ with respect to $\lambda > 0$ yields

$$\begin{aligned}
&(1-q_{t-1})\pi(\omega^R)[-(1-\pi(\omega^B))(1-F_X(k)) + (1-\pi(\omega^B))(1-F_Y(k))]F_X(\widehat{k}_t^\lambda(I|\emptyset)) \\
&+(1-q_{t-1})\pi(\omega^R)[-(1-\pi(\omega^B))F_X(k) + (1-\pi(\omega^B))F_Y(k)]F_X(\widehat{k}_t^\lambda(N|\emptyset)) \\
&+(1-q_{t-1})(1-\pi(\omega^R))[\pi(\omega^B))(1-F_X(k)) - \pi(\omega^B)(1-F_Y(k))]F_Y(\widehat{k}_t^\lambda(I|\emptyset)) \\
&+(1-q_{t-1})(1-\pi(\omega^R))[\pi(\omega^B)F_X(k) - \pi(\omega^B)F_Y(k)]F_Y(\widehat{k}_t^\lambda(N|\emptyset)),
\end{aligned}$$

which can be rewritten as

$$\begin{aligned}&(1-q_{t-1})\pi(\omega^R)(1-\pi(\omega^B))[F_Y(k)-F_X(k)][F_X(\widehat{k}_t^\lambda(N|\emptyset))-F_X(\widehat{k}_t^\lambda(I|\emptyset))]\\&-(1-q_{t-1})(1-\pi(\omega^R))\pi(\omega^B)[F_Y(k)-F_X(k)][F_Y(\widehat{k}_t^\lambda(N|\emptyset))-F_Y(\widehat{k}_t^\lambda(I|\emptyset))].\end{aligned} \tag{44}$$

Let's examine the above equation (44) closely for each $\omega^A \in \Omega^A$. By substituting (3) into it, the following four cases are derived.
(i) $\omega^A = I|\emptyset$

$$\frac{(1-q_{t-1})[F_Y(k)-F_X(k)]}{2(2-F_X(k)-F_Y(k))}\begin{Bmatrix}(1-F_X(k))[F_X(\widehat{k}_t^\lambda(N|\emptyset))-F_X(\widehat{k}_t^\lambda(I|\emptyset))]\\-(1-F_Y(k))[F_Y(\widehat{k}_t^\lambda(N|\emptyset))-F_Y(\widehat{k}_t^\lambda(I|\emptyset))]\end{Bmatrix},$$

(ii) $\omega^A = \emptyset|I$

$$\frac{(1-q_{t-1})[F_Y(k)-F_X(k)]}{2(2-F_X(k)-F_Y(k))}\begin{Bmatrix}(1-F_Y(k))[F_X(\widehat{k}_t^\lambda(N|\emptyset))-F_X(\widehat{k}_t^\lambda(I|\emptyset))]\\-(1-F_X(k))[F_Y(\widehat{k}_t^\lambda(N|\emptyset))-F_Y(\widehat{k}_t^\lambda(I|\emptyset))]\end{Bmatrix},$$

(iii) $\omega^A = \emptyset|N$

$$\frac{(1-q_{t-1})[F_Y(k)-F_X(k)]}{2(F_X(k)+F_Y(k))}\begin{Bmatrix}F_Y(k)[F_X(\widehat{k}_t^\lambda(N|\emptyset))-F_X(\widehat{k}_t^\lambda(I|\emptyset))]\\-F_X(k)[F_Y(\widehat{k}_t^\lambda(N|\emptyset))-F_Y(\widehat{k}_t^\lambda(I|\emptyset))]\end{Bmatrix},$$

(iv) $\omega^A = N|\emptyset$

$$\frac{(1-q_{t-1})[F_Y(k)-F_X(k)]}{2(F_X(k)+F_Y(k))}\begin{Bmatrix}F_X(k)[F_X(\widehat{k}_t^\lambda(N|\emptyset))-F_X(\widehat{k}_t^\lambda(I|\emptyset))]\\-F_Y(k)[F_Y(\widehat{k}_t^\lambda(N|\emptyset))-F_Y(\widehat{k}_t^\lambda(I|\emptyset))]\end{Bmatrix}.$$

Then, by $\widehat{k}_t^\lambda(N|\emptyset) > \widehat{k}_t^\lambda(I|\emptyset)$ from Step 1, if (A6) holds, we have a positive sign for all four cases.

Part 2. We show (ii) for q_{t-1} sufficiently close to 1, since $q_{t-1} \to 1$ as $\Delta\mathcal{P}_{t-1} \to 1$ from (4). Note that $\widehat{\Phi}_t(\widehat{k}_t^\lambda, \omega^A, \lambda)$ is continuously differentiable in λ, q_{t-1}, so for each $\omega^A \in \Omega^A$, we have

$$\frac{\partial^2 \widehat{\Phi}_t(\widehat{k}_t^\lambda, \omega^A, \lambda)}{\partial\lambda\partial q_{t-1}} < 0,$$

which implies

$$q_{t-1} \to 1, \frac{\partial \widehat{\Phi}_t(\widehat{k}_t^\lambda, \omega^A, \lambda)}{\partial\lambda} \to 0 \text{ for each } \lambda > 0 \text{ and } q_{t-1} > \frac{1}{2}.$$

This establishes that a homogeneous mapping in (37) and a heterogeneous mapping in (41) are uniformly close to each other for a sufficiently high q_{t-1}. ■

Proof of Lemma 3. For the lemma, we suppose $\alpha = 1$.

(i) We show $\ell_t^B(\omega) = \ell_t^R(\omega)$ for all $\omega \in \Omega$. First, by combining the location decision in (22) and Proposition 4, a location equilibrium as given by Definition 1 can be simplified such that ℓ_t is a location equilibrium if (i) for each $\omega \in \Omega$,

$$\ell_t^B(\omega) = \begin{cases} E & \text{if } \Delta P_t(\omega, \ell_t)[\mathbf{1}_{\{\omega \in \Omega^{B+}\}} - \mathbf{1}_{\{\omega \in \Omega^{B-}\}}] > 0, \\ W & \text{if } \Delta P_t(\omega, \ell_t)[\mathbf{1}_{\{\omega \in \Omega^{B+}\}} - \mathbf{1}_{\{\omega \in \Omega^{B-}\}}] < 0, \end{cases}$$

$$\ell_t^R(\omega) = \begin{cases} E & \text{if } \Delta P_t(\omega, \ell_t)[\mathbf{1}_{\{\omega \in \Omega^{R+}\}} - \mathbf{1}_{\{\omega \in \Omega^{R-}\}}] < 0, \\ W & \text{if } \Delta P_t(\omega, \ell_t)[\mathbf{1}_{\{\omega \in \Omega^{R+}\}} - \mathbf{1}_{\{\omega \in \Omega^{R-}\}}] > 0, \end{cases}$$

and (ii) $P_t^B(\omega, \ell_t^B) = \mathbb{E}\big[\mathbf{1}_{\{\ell_t^B(\tilde{\omega})=E\}} \mid \omega\big]$ and $P_t^R(\omega, \ell_t^R) = \mathbb{E}\big[\mathbf{1}_{\{\ell_t^R(\tilde{\omega})=E\}} \mid \omega\big]$. By $\Omega^{B+} = \Omega^{R-}$ and $\Omega^{B-} = \Omega^{R+}$ in (10), $\mathbf{1}_{\{\omega \in \Omega^{B+}\}} - \mathbf{1}_{\{\omega \in \Omega^{B-}\}} = -[\mathbf{1}_{\{\omega \in \Omega^{R+}\}} - \mathbf{1}_{\{\omega \in \Omega^{R-}\}}]$ for all ω, so we have $\ell_t^B(\omega) = \ell_t^R(\omega)$ for $\Delta P_t(\omega, \ell_t) \neq 0$.

(ii) Given $\ell_t^B(\omega) = \ell_t^R(\omega)$, by combining (32) and (33), we have

$$\begin{aligned} \Delta P_t(\omega, \ell_t) &= \mathbb{E}\big[\mathbf{1}_{\{\ell_t^B(\tilde{\omega})=E\}} \mid \omega\big] - \mathbb{E}\big[\mathbf{1}_{\{\ell_t^R(\tilde{\omega})=E\}} \mid \omega\big] \\ &= (2q_{t-1} - 1)p_I(\omega^B)\mathbf{1}_{\{\ell_t^g(I,\emptyset)=E\}} - (2q_{t-1} - 1)p_N(\omega^R)\mathbf{1}_{\{\ell_t^g(\emptyset,N)=E\}} \\ &\quad + (2q_{t-1} - 1)p_N(\omega^B)\mathbf{1}_{\{\ell_t^g(N,\emptyset)=E\}} - (2q_{t-1} - 1)p_I(\omega^R)\mathbf{1}_{\{\ell_t^g(\emptyset,I)=E\}}, \end{aligned}$$

which results in (45) given $p_I(\omega^R) = 1 - p_N(\omega^R)$ and $p_I(\omega^B) = 1 - p_N(\omega^B)$:

$$\begin{aligned} \Delta P_t(\omega, \ell_t) =& (2q_{t-1} - 1)[\mathbf{1}_{\{\ell_t^g(I,\emptyset)=E\}} - \mathbf{1}_{\{\ell_t^g(\emptyset,I)=E\}}] \\ &+ (2q_{t-1} - 1)p_N(\omega^B)[\mathbf{1}_{\{\ell_t^g(N,\emptyset)=E\}} - \mathbf{1}_{\{\ell_t^g(I,\emptyset)=E\}}] \\ &+ (2q_{t-1} - 1)p_N(\omega^R)[\mathbf{1}_{\{\ell_t^g(\emptyset,I)=E\}} - \mathbf{1}_{\{\ell_t^g(\emptyset,N)=E\}}]. \end{aligned} \tag{45}$$

We divide the proof into two parts with four cases.

Part 1. Consider Ω^{B+} and note that for each $\omega = (\omega^B, \omega^R) \in \Omega^{B+}$, $p_N(\omega^B) < p_N(\omega^R)$ from (5). Suppose $\mathbf{1}_{\{\ell_t(I,\emptyset)=E\}} \neq \mathbf{1}_{\{\ell_t(\emptyset,N)=E\}}$. *Case 1.* First, suppose $\mathbf{1}_{\{\ell_t(I,\emptyset)=E\}} = 1$, $\mathbf{1}_{\{\ell_t(\emptyset,N)=E\}} = 0$. Since $(\emptyset, N) \in \Omega^{B+}$ but $\mathbf{1}_{\{\ell_t(\emptyset,N)=E\}} = 0$ (choosing W), from (45), for $\omega = (\emptyset, N)$, there must be more B players in W, that is,

$$\begin{aligned} 0 &> \Delta P_t((\emptyset, N), \ell_t) \\ &= (2q_{t-1} - 1)\left[(1 - p_N(\emptyset^B)) - (1 - p_N(N^R))\mathbf{1}_{\{\ell_t(\emptyset,I)=E\}} + p_N(\emptyset^B)\mathbf{1}_{\{\ell_t(N,\emptyset)=E\}}\right] \\ &> (2q_{t-1} - 1)\left[(1 - p_N(N^R))(1 - \mathbf{1}_{\{\ell_t(\emptyset,I)=E\}}) + p_N(\emptyset^B)\mathbf{1}_{\{\ell_t(N,\emptyset)=E\}}\right], \end{aligned}$$

where the last inequality follows from $1 - p_N(\emptyset^B) > 1 - p_N(N^R)$. However, $(1 - p_N(N^R))(1 - \mathbf{1}_{\{\ell_t(\emptyset,I)=E\}}) + p_N(\emptyset^B)\mathbf{1}_{\{\ell_t(N,\emptyset)=E\}} \geq 0$ for all $\mathbf{1}_{\{\ell_t(N,\emptyset)=E\}}, \mathbf{1}_{\{\ell_t(\emptyset,I)=E\}}$. We have a contradiction. *Case 2.* Now, suppose $\mathbf{1}_{\{\ell_t(I,\emptyset)=E\}} = 0$, $\mathbf{1}_{\{\ell_t(\emptyset,N)=E\}} = 1$. Since $(\emptyset, N) \in \Omega^{B+}$ but $\mathbf{1}_{\{\ell_t(\emptyset,N)=E\}} = 1$ (choosing E), from (45), for $\omega = (\emptyset, N)$, there must be more B players in E, that is,

$$\begin{aligned} 0 &< \Delta P_t((\emptyset, N), \ell_t) \\ &= (2q_{t-1} - 1)\left[-p_N(N^R) - (1 - p_N(N^R))\mathbf{1}_{\{\ell_t(\emptyset,I)=E\}} + p_N(\emptyset^B)\mathbf{1}_{\{\ell_t(N,\emptyset)=E\}}\right] \\ &< (2q_{t-1} - 1)\left[-p_N(\emptyset^B)(1 - \mathbf{1}_{\{\ell_t(N,\emptyset)=E\}}) - (1 - p_N(N^R))\mathbf{1}_{\{\ell_t(\emptyset,I)=E\}}\right], \end{aligned}$$

where the last inequality follows from $-p_N(\emptyset^B) > -p_N(N^R)$. However, $-p_N(\emptyset^B)(1 - \mathbf{1}_{\{\ell_t(N,\emptyset)=E\}}) - (1 - p_N(N^R))\mathbf{1}_{\{\ell_t(\emptyset,I)=E\}} \leq 0$ for all $\mathbf{1}_{\{\ell_t(N,\emptyset)=E\}}, \mathbf{1}_{\{\ell_t(\emptyset,I)=E\}}$. We have a contradiction.

Part 2. Consider Ω^{B-} and note that for each $\omega = (\omega^B, \omega^R) \in \Omega^{B+}$, $p_N(\omega^B) > p_N(\omega^R)$ from (5). Suppose $\mathbf{1}_{\{\ell_t(N,\emptyset)=E\}} \neq \mathbf{1}_{\{\ell_t(\emptyset,I)=E\}}$. *Case 1.* First, suppose $\mathbf{1}_{\{\ell_t(N,\emptyset)=E\}} = 1$, $\mathbf{1}_{\{\ell_t(\emptyset,I)=E\}} = 0$. Since $(N,\emptyset) \in \Omega^{B-}$ but $\mathbf{1}_{\{\ell_t(N,\emptyset)=E\}} = 1$ (choosing E), from (45), for $\omega = (N,\emptyset)$, there must be more B players in W, that is,

$$\begin{aligned}
0 &> \Delta P_t((N,\emptyset),\ell_t) \\
&= (2q_{t-1} - 1)\left[p_N(N^B) + (1 - p_N(N^B))\mathbf{1}_{\{\ell_t(I,\emptyset)=E\}} - p_N(\emptyset^R)\mathbf{1}_{\{\ell_t(\emptyset,N)=E\}}\right] \\
&> (2q_{t-1} - 1)\left[p_N(\emptyset^R)(1 - \mathbf{1}_{\{\ell_t(\emptyset,I)=E\}}) + (1 - p_N(N^B))\mathbf{1}_{\{\ell_t(I,\emptyset)=E\}}\right],
\end{aligned}$$

where the last inequality follows from $p_N(N^B) > p_N(\emptyset^R)$. However, $p_N(\emptyset^R)(1-\mathbf{1}_{\{\ell_t(\emptyset,I)=E\}}) + (1 - p_N(N^B))\mathbf{1}_{\{\ell_t(I,\emptyset)=E\}} \geq 0$ for all $\mathbf{1}_{\{\ell_t(I,\emptyset)=E\}}, \mathbf{1}_{\{\ell_t(\emptyset,N)=E\}}$. We have a contradiction. *Case 2.* Now, suppose $\mathbf{1}_{\{\ell_t(N,\emptyset)=E\}} = 0$, $\mathbf{1}_{\{\ell_t(\emptyset,I)=E\}} = 1$. Since $(N,\emptyset) \in \Omega^{B-}$ but $\mathbf{1}_{\{\ell_t(N,\emptyset)=E\}} = 0$ (choosing W), from (45), for $\omega = (N,\emptyset)$, there must be more B players in E, that is,

$$\begin{aligned}
0 &< \Delta P_t((N,\emptyset),\ell_t) \\
&= (2q_{t-1} - 1)\left[-(1 - p_N(\emptyset^R)) + (1 - p_N(N^B))\mathbf{1}_{\{\ell_t(I,\emptyset)=E\}} - p_N(\emptyset^R)\mathbf{1}_{\{\ell_t(\emptyset,N)=E\}}\right] \\
&< (2q_{t-1} - 1)\left[-(1 - p_N(N^B))(1 - \mathbf{1}_{\{\ell_t(I,\emptyset)=E\}}) - p_N(\emptyset^R)\mathbf{1}_{\{\ell_t(\emptyset,N)=E\}}\right],
\end{aligned}$$

where the last inequality follows from $p_N(N^B) > p_N(\emptyset^R)$. However, $-(1 - p_N(N^B))(1 - \mathbf{1}_{\{\ell_t(I,\emptyset)=E\}}) - p_N(\emptyset^R)\mathbf{1}_{\{\ell_t(\emptyset,N)=E\}} \leq 0$ for all $\mathbf{1}_{\{\ell_t(I,\emptyset)=E\}}, \mathbf{1}_{\{\ell_t(\emptyset,N)=E\}}$. We have a contradiction. ■

Proof of Proposition 5. *Step 1.* We show that (i) of Lemma 3 holds for $\alpha \in (0,1)$. Recall in the model (Section 2) that an optimal location strategy ℓ_t is robust if there exists a sufficiently small open interval I around α such that the strategy ℓ_t is optimal for all $\alpha' \in I$. We first show that any location strategy with $\Delta P_t(\omega,\ell_t) = 0$ but $\Delta\mathcal{P}_{t-1} \neq 0$ for some $\omega \in \Omega$ is not robust. Suppose there exists a location strategy $\ell_t = (\ell_t^B, \ell_t^R)$ with $\Delta P_t(\omega,\ell_t) = 0$ and $\Delta\mathcal{P}_{t-1} \neq 0$ for a history $\omega \in \Omega$. By incorporating a location strategy $\ell_t^g : \Omega \times [-1,1] \to L$ into (32) and (33) with $\alpha > 0$, for the history $\omega \in \Omega$, we have

$$0 = \Delta P_t(\omega,\ell_t) = \alpha\left\{\mathbb{E}\left[\mathbf{1}_{\{\ell_t^B(\widetilde{\omega},z_{t-1})=E\}} \mid \omega\right] - \mathbb{E}\left[\mathbf{1}_{\{\ell_t^B(\widetilde{\omega},z_{t-1})=E\}} \mid \omega\right]\right\} + (1-\alpha)\Delta\mathcal{P}_{t-1}.$$

This means that for some $\Delta\mathcal{P}_{t-1} \neq 0$, there exist $\ell_t^B : \Omega\times[-1,1] \to L$ and $\ell_t^R : \Omega\times[-1,1] \to L$ satisfying the equality above. Note that this holds for a particular α value, so it is not robust. This implies that for a robust location strategy, $\Delta P_t(\omega,\ell_t) = 0$ if and only if $\Delta\mathcal{P}_{t-1} = 0$.

Step 2. We show that (ii) of Lemma 3 holds for $\alpha \in (0,1)$. Without loss of generality, examine $\Delta\mathcal{P}_{t-1} > 0$. Consider Part 1 in the proof of Lemma 3 now with $\alpha \in (0,1)$. Given the sign $\Delta\mathcal{P}_{t-1} > 0$, Case 1 of Part 1 still yields a contradiction. To examine the other case,

Case 2 of Part 1, rewrite it with $\alpha \in (0,1)$ such that

$$\begin{aligned}
0 &< \Delta P_t((\emptyset, N), \ell_t) \\
&= \alpha(2q_{t-1}-1)\left[-p_N(N^R) - (1-p_N(N^R))\mathbf{1}_{\{\ell_t((\emptyset,I),z_{t-1})=E\}} + p_N(\emptyset^B)\mathbf{1}_{\{\ell_t((N,\emptyset),z_{t-1})=E\}}\right] \\
&\quad + (1-\alpha)\Delta\mathcal{P}_{t-1} \\
&< \alpha(2q_{t-1}-1)\left[-p_N(\emptyset^B)(1-\mathbf{1}_{\{\ell_t((N,\emptyset),z_{t-1})=E\}}) - (1-p_N(N^R))\mathbf{1}_{\{\ell_t((\emptyset,I),z_{t-1})=E\}}\right] + (1-\alpha)\Delta\mathcal{P}_{t-1}.
\end{aligned}$$

Then, for a sufficiently large $\Delta\mathcal{P}_{t-1} > 0$, there is no contradiction. This implies that $\mathbf{1}_{\{\ell_t((I,\emptyset),z_{t-1})=E\}} = 0$, $\mathbf{1}_{\{\ell_t((\emptyset,N),z_{t-1})=E\}} = 1$ can be a part of an equilibrium location strategy. In order for this choice to be a part of an equilibrium location strategy, we need $\Delta P_t((\emptyset, N), \ell_t) > 0$ in (45), and given $\mathbf{1}_{\{\ell_t((I,\emptyset),z_{t-1})=E\}} = 0$, $\mathbf{1}_{\{\ell_t((\emptyset,N),z_{t-1})=E\}} = 1$, $\Delta P_t((\emptyset, N), \ell_t) > 0$ requires $\mathbf{1}_{\{\ell_t((N,\emptyset),z_{t-1})=E\}} = 1$. Now, Case 1 of Part 2 implies that if $\mathbf{1}_{\{\ell_t((N,\emptyset),z_{t-1})=E\}} = 1$, then $\mathbf{1}_{\{\ell_t((\emptyset,I),z_{t-1})=E\}} = 1$ as well. But this yields $\Delta P_t((\emptyset, N), \ell_t) < 0$ in (45), a contradiction.

Similarly, consider Part 2 in the proof of Lemma 3 now with $\alpha \in (0,1)$. Given the sign $\Delta\mathcal{P}_{t-1} > 0$, Case 1 of Part 2 still yields a contradiction. To examine the other case, Case 2 of Part 2, rewrite it with $\alpha \in (0,1)$ such that

$$\begin{aligned}
0 &< \Delta P_t((N, \emptyset), \ell_t) \\
&= \alpha(2q_{t-1}-1)\left[-(1-p_N(\emptyset^R)) + (1-p_N(N^B))\mathbf{1}_{\{\ell_t((I,\emptyset),z_{t-1})=E\}} - p_N(\emptyset^R)\mathbf{1}_{\{\ell_t((\emptyset,N),z_{t-1})=E\}}\right] \\
&\quad + (1-\alpha)\Delta\mathcal{P}_{t-1} \\
&< \alpha(2q_{t-1}-1)\left[-(1-p_N(N^B))(1-\mathbf{1}_{\{\ell_t((I,\emptyset),z_{t-1})=E\}}) - p_N(\emptyset^R)\mathbf{1}_{\{\ell_t((\emptyset,N),z_{t-1})=E\}}\right] + (1-\alpha)\Delta\mathcal{P}_{t-1},
\end{aligned}$$

Then, for a sufficiently large $\Delta\mathcal{P}_{t-1} > 0$, there is no contradiction. This implies that $\mathbf{1}_{\{\ell_t((N,\emptyset),z_{t-1})=E\}} = 0$, $\mathbf{1}_{\{\ell_t((\emptyset,I),z_{t-1})=E\}} = 1$. can be a part of an equilibrium location strategy. With a similar procedure as above, now, Case 1 of Part 1 implies that $\mathbf{1}_{\{\ell_t((I,\emptyset),z_{t-1})=E\}} = 1$, $\mathbf{1}_{\{\ell_t((\emptyset,N),z_{t-1})=E\}} = 1$ as well. But this yields $\Delta P_t((\emptyset, N), \ell_t) < 0$ in (45), a contradiction.

Step 3. By Steps 1-2, extending Lemma 3, if there is any robust location equilibrium strategy, then it is a binary splitting location equilibrium in Table 2 (again, a strategy with all players moving to one location is not an equilibrium with $\alpha \in (0,1)$). Hence, the statement in Proposition 5 can be rewritten with the existence of a binary splitting location equilibrium.

Part 1. We first show that there exists a binary splitting location equilibrium in Table 2 if and only if $1 - \widehat{p}_N(N, \emptyset) > 0$. Consider the formula in (24). "If part" can be drawn from the one in the proof of Proposition 1, including $1 - \widehat{p}_N(\omega) > 0$ for all $\omega \in \Omega$ if and only if $1 - \widehat{p}_N(N, \emptyset) > 0$, so if $1 - \widehat{p}_N(N, \emptyset) > 0$, then there exists a binary splitting location equilibrium. Now, suppose a binary splitting location equilibrium and $1 - \widehat{p}_N(\omega) \leq 0$ for some $\omega \in \Omega$. If $1 - \widehat{p}_N(\omega) < 0$, we have a contradiction with the binary splitting location strategy in Table 2 for α sufficiently close to 1 since the system in (24) is sufficiently close to the system in (14), and if $1 - \widehat{p}_N(\omega) = 0$, we have a contradiction with the first branch of the formula in (24) if $\Delta\mathcal{P}_{t-1} < 0$ and the second branch of the formula in (24) if $\Delta\mathcal{P}_{t-1} > 0$.

Part 2. Now, we show that any robust location strategy is a Markov location strategy in (25). For this, we just show that for $\Delta\mathcal{P}_{t-1} > 0$, an equilibrium with $\Delta P_t(\omega, \ell_t) > 0$ is only

robust; the other case can be symmetrically shown. Consider $\Delta\mathcal{P}_{t-1} > 0$. For $\Delta P_t(\omega, \ell_t) < 0$, the square function yields a positive horizontal intercept to make (24) zero (see the right panel of Figure 5), which is given as $\frac{(1-\alpha)\gamma}{\alpha}\frac{A_{t-1}(2-A_{t-1})}{1-\widehat{p}_N(\omega)}$. As in part 1, it is without loss of generality for us to choose $\omega = (N, \emptyset)$ for $\widehat{p}_N(\omega)$ instead of all ω. Suppose $\Delta\mathcal{P}_{t-1} > \frac{(1-\alpha)\gamma}{\alpha}\frac{A_{t-1}(2-A_{t-1})}{1-\widehat{p}_N((N,\emptyset))}$. Then, both $\Delta P_t(\omega, \ell_t) > 0$ and $\Delta P_t(\omega, \ell_t) < 0$ equilibria can arise. But the latter equilibrium $\Delta P_t(\omega, \ell_t) < 0$ is not robust if we consider now $\Delta\mathcal{P}_{t-1} < \frac{(1-\alpha)\gamma}{\alpha}\frac{A_{t-1}(2-A_{t-1})}{1-\widehat{p}_N((N,\emptyset))}$. An equilibrium with $\Delta P_t(\omega, \ell_t) < 0$ cannot arise since the square function yields a positive value for $\omega = (N, \emptyset)$, contradicting $\Delta P_t(\omega, \ell_t) < 0$. Hence, $\frac{(1-\alpha)\gamma}{\alpha}\frac{A_{t-1}(2-A_{t-1})}{1-\widehat{p}_N((N,\emptyset))}$ depends on a particular value of α, so the only robust equilibrium belief is that $\Delta\mathcal{P}_{t-1} > 0\,(< 0)$ yields $\Delta P_t(\omega, \ell_t) > 0\,(< 0)$, which is corresponding to a Markov location strategy in (25).

With a Markov location strategy in (25), the belief dynamics (24) changes into

$$\Delta P_t(\omega, \ell_t) = \begin{cases} \alpha\frac{1-\widehat{p}_N(\omega)}{A_{t-1}(2-A_{t-1})}\Delta\mathcal{P}_{t-1}^2 + (1-\alpha)\Delta\mathcal{P}_{t-1} & \text{if } \Delta\mathcal{P}_{t-1} \geq 0, \\ -\alpha\frac{1-\widehat{p}_N(\omega)}{A_{t-1}(2-A_{t-1})}\Delta\mathcal{P}_{t-1}^2 + (1-\alpha)\Delta\mathcal{P}_{t-1} & \text{if } \Delta\mathcal{P}_{t-1} \leq 0, \end{cases} \tag{46}$$

where note that $\Delta P_t(\omega, \ell_t)$ only depends on the sign of $\Delta\mathcal{P}_{t-1}$, unlike (24). ■

Proof of Lemma 4. From an outside theorist's point of view, knowing the true distributions, $p_N(\omega^B) = F_B(k)$ and $p_N(\omega^R) = F_R(k)$, and the summation of them yields (29). Then, one can show that A_t converges to $1 - F_B(k) + F_R(k)$ such that $A_t > A_{t-1}$ for all $t = 1, 2, ...$ if $A_0 < 1 - F_B(k) + F_R(k)$, whereas $A_t < A_{t-1}$ for all $t = 1, 2, ...$ if $A_0 > 1 - F_B(k) + F_R(k)$ since from (29), we have

$$A_t - A_{t-1} = \begin{cases} \alpha[1 - F_B(k) + F_R(k) - A_{t-1}] > 0 & \text{if } A_0 < 1 - F_B(k) + F_R(k), \\ \alpha[1 - F_B(k) + F_R(k) - A_{t-1}] < 0 & \text{if } A_0 > 1 - F_B(k) + F_R(k). \end{cases}$$

For any combination of (F_B, F_R), if $A_0 < 1 - |F_B(k) - F_R(k)|$, A_{t-1} strictly increases, whereas $A_0 > 1 + |F_B(k) - F_R(k)|$, A_{t-1} strictly decreases. Both cases make $A_{t-1}(2 - A_{t-1})$ increase, so z_{t-1}^* through (28). ■

Proof of Proposition 6. First, consider the binary splitting location strategy in Table 2. By incorporating the exogenous moves into (32) and (33), we have[38]

$$P_t^B(\omega, \ell_t^B) = \begin{cases} \alpha[q_{t-1}p_I(\omega^B) + (1-q_{t-1})p_N(\omega^R)] + (1-\alpha)(\mathcal{P}_{t-1}^B + \epsilon\Delta\mathcal{P}_{t-1}) & \text{if } \Delta P_t(\omega, \ell_t) > 0, \\ \alpha[q_{t-1}p_N(\omega^B) + (1-q_{t-1})p_I(\omega^R)] + (1-\alpha)(\mathcal{P}_{t-1}^B + \epsilon\Delta\mathcal{P}_{t-1}) & \text{if } \Delta P_t(\omega, \ell_t) < 0, \end{cases}$$

and

$$P_t^R(\omega, \ell_t^R) = \begin{cases} \alpha[(1-q_{t-1})p_I(\omega^B) + q_{t-1}p_N(\omega^R)] + (1-\alpha)(\mathcal{P}_{t-1}^R - \epsilon\Delta\mathcal{P}_{t-1}) & \text{if } \Delta P_t(\omega, \ell_t) > 0, \\ \alpha[(1-q_{t-1})p_N(\omega^B) + q_{t-1}p_I(\omega^R)] + (1-\alpha)(\mathcal{P}_{t-1}^R - \epsilon\Delta\mathcal{P}_{t-1}) & \text{if } \Delta P_t(\omega, \ell_t) < 0. \end{cases}$$

[38]With systematic polarization, more precisely, $P_t^B(\omega, \ell_t^B)$ is the minimum of the expression on the RHS and 1.

Then with a Markov strategy in (25), this yields beliefs dynamics such that

$$\Delta P_t(\omega,\ell_t)=\begin{cases}\alpha\frac{1-\widehat{p}_N(\omega)}{A_{t-1}(2-A_{t-1})}\Delta\mathcal{P}_{t-1}^2+(1-\alpha)\gamma\Delta\mathcal{P}_{t-1} & \text{if } \Delta\mathcal{P}_{t-1}\geq 0,\\ -\alpha\frac{1-\widehat{p}_N(\omega)}{A_{t-1}(2-A_{t-1})}\Delta\mathcal{P}_{t-1}^2+(1-\alpha)\gamma\Delta\mathcal{P}_{t-1} & \text{if } \Delta\mathcal{P}_{t-1}\leq 0,\end{cases}$$

and real dynamics in (26).

Now, the difference between $\Delta\mathcal{P}_t$ and $\Delta\mathcal{P}_{t-1}$ is given as

$$\Delta\mathcal{P}_t-\Delta\mathcal{P}_{t-1}=\begin{cases}\big[\alpha\beta_{t-1}\Delta\mathcal{P}_{t-1}+(1-\alpha)\gamma-1\big]\Delta\mathcal{P}_{t-1} & \text{if } \Delta\mathcal{P}_{t-1}\geq 0,\\ \big[-\alpha\beta_{t-1}\Delta\mathcal{P}_{t-1}+(1-\alpha)\gamma-1\big]\Delta\mathcal{P}_{t-1} & \text{if } \Delta\mathcal{P}_{t-1}\leq 0.\end{cases}$$

Then, given $\alpha\beta_{\tau-1}\Delta\mathcal{P}_{\tau-1}+(1-\alpha)\gamma-1\leq 0$, we have $\Delta\mathcal{P}_\tau\leq\Delta\mathcal{P}_{\tau-1}$ for $\tau=1,2,....$ If A_0 satisfies Lemma 4, then $z_t^*>z_{t-1}^*$ for all $t=1,2,....$, which implies that $\beta_t<\beta_{t-1}$. With it, we show that if $\Delta\mathcal{P}_t-\Delta\mathcal{P}_{t-1}\leq 0$, then $\Delta\mathcal{P}_{t+1}-\Delta\mathcal{P}_t<0$ for all $t\geq 0$ such that

$$\begin{aligned}\Delta\mathcal{P}_{t+1}-\Delta\mathcal{P}_t&=\big[\alpha\beta_t\Delta\mathcal{P}_t+(1-\alpha)\gamma-1\big]\Delta\mathcal{P}_t\\&<\big[\alpha\beta_{t-1}\Delta\mathcal{P}_{t-1}+(1-\alpha)\gamma-1\big]\Delta\mathcal{P}_t\\&=\left[\frac{\Delta\mathcal{P}_t-\Delta\mathcal{P}_{t-1}}{\Delta\mathcal{P}_{t-1}}\right]\Delta\mathcal{P}_t.\end{aligned}$$

Hence, $\Delta\mathcal{P}_{t+1}<\Delta\mathcal{P}_t$ for all $t\geq\tau$. ■

Proof of Proposition 7. First, we show that if for each $t=1,2,....$,

$$\alpha\beta_t\Delta\mathcal{P}_t+(1-\alpha)\gamma\geq 1\Rightarrow\alpha\beta_{t-1}\Delta\mathcal{P}_{t-1}+(1-\alpha)\gamma>1. \tag{47}$$

Suppose, on the contrary, that there exists t such that $\alpha\beta_{t-1}\Delta\mathcal{P}_{t-1}+(1-\alpha)\gamma\leq 1$ and $\alpha\beta_t\Delta\mathcal{P}_t+(1-\alpha)\gamma\geq 1$. Then, $\Delta\mathcal{P}_t\leq\Delta\mathcal{P}_{t-1}$. In addition, $\beta_t<\beta_{t-1}$ by the proof of Lemma 4. Hence,

$$\alpha\beta_{t-1}\Delta\mathcal{P}_{t-1}+(1-\alpha)\gamma\geq\alpha\beta_{t-1}\Delta\mathcal{P}_t+(1-\alpha)\gamma>\alpha\beta_t\Delta\mathcal{P}_t+(1-\alpha)\gamma\geq 1,$$

which is a contradiction. Now, for the proof, we define f_{t-1} in (26), separately, as the one for the positive domain and that for the negative domain, respectively, such as

$$\begin{aligned}&f_{t-1}^{(+)}(z)\equiv\alpha\beta_{t-1}z^2+(1-\alpha)\gamma z \text{ if } z>0,\\&f_{t-1}^{(-)}(z)\equiv-\alpha\beta_{t-1}z^2+(1-\alpha)\gamma z \text{ if } z<0.\end{aligned} \tag{48}$$

and their inverse functions as $h_{t-1}^{(+)}(z)$ and $h_{t-1}^{(-)}(z)$. In addition, we denote their fixed points by $z_t^{(+)}$ and $z_t^{(-)}$.

Part (i) $1-F_N>0$. Find $\widehat{t}$ such that $\widehat{t}\equiv\max\{t:z_t^{(+)}\leq 1\}$. Then, given $\widehat{t}$, we obtain $\Delta\mathcal{P}_{\widehat{t}}=h_{\widehat{t}}^{(+)}(1)$. This implies that for $\Delta\mathcal{P}_{\widehat{t}}=h_{\widehat{t}}^{(+)}(1)$, $\alpha\beta_t\Delta\mathcal{P}_{\widehat{t}}+(1-\alpha)\gamma\geq 1$, so by (47), for each $t<\widehat{t}$,

$$\alpha\beta_t\Delta\mathcal{P}_t+(1-\alpha)\gamma>1.$$

To find the minimal initial difference, by the sequence of inverse functions of $f_{t-1}^{(+)}(z)$ in (48),

$$z' \equiv h_0^{(+)}\Big(h_1^{(+)}\big(\cdots h_{\widehat{t}}^{(+)}(1)\cdots\big)\Big).$$

Finally, $z^{\dagger} = \max\{z', \Delta\overline{\mathcal{P}}_0\}$, where recall that $\Delta\overline{\mathcal{P}}_0$ is a critical level satisfying a relatively high difference in Proposition 4.

Part (ii) Consider a sequence of functions such that $f_0^{(+)}, f_1^{(-)}, ...$ with their corresponding fixed points such as $z_0^{(+)}, z_1^{(-)}, ...$ Now, for $\tau = 1, 2, ...,$ denote

$$z_t^a = \begin{cases} z_t^{(+)} & \text{if } t = 2\tau - 2, \\ z_t^{(-)} & \text{if } t = 2\tau - 1. \end{cases}$$

Find $\widehat{t}$ such that $\widehat{t} \equiv \max\Big\{t : \big|z_t^a\big| \leq 1\Big\}$. Then, given $\widehat{t}$, we obtain $\Delta\mathcal{P}_{\widehat{t}} = h_{\widehat{t}}^{(s)}(1)$, where

$$s = \begin{cases} + & \text{if } f_{\widehat{t}}^{(+)}(\Delta\mathcal{P}_{\widehat{t}}) = -1, \\ - & \text{if } f_{\widehat{t}}^{(-)}(\Delta\mathcal{P}_{\widehat{t}}) = 1. \end{cases}$$

Then, using a similar procedure to Part (i), we find

$$z'' \equiv h_0^{(+)}\Big(h_1^{(-)}\big(\cdots h_{\widehat{t}}^{(s)}(1)\cdots\big)\Big).$$

As in Part (i), $z^{\dagger\dagger} = \max\{z'', \Delta\overline{\mathcal{P}}_0\}$. This completes the proof. ■

Proof of Proposition 8. We show only the case where $A_0 > 1$ and divide the proof into two parts.

Case 1. $F_B < F_R$. Denote the sum of the moving probabilities with $F_B < F_R$ by A'_{t-1} and the sum with $F_B = F_R$ by A_{t-1}. Since by Lemma 4, $A_{t-1} > 1$, we have $\alpha + (1-\alpha)A_{t-1} > 1$. Then, given $1 - F_B(k) + F_R(k) > 1$, by comparing $F_B < F_R$ with $F_B = F_R$,

$$\alpha[1 - F_B(k) + F_R(k)] + (1-\alpha)A'_{t-1} > \alpha + (1-\alpha)A_{t-1},$$

which implies that $A'_{t-1}(2 - A'_{t-1}) < A_{t-1}(2 - A_{t-1})$ for all t since the function $a(2-a)$ is strictly decreasing in $a > 1$. Further, denote $F'_N = F_X(k) + F_Y(k)$ for the different distributions and $F_N = F_Y(k) + F_Y(k)$ for the same distribution. By FOSD between F_X and F_Y, we have $1 - F'_N > 1 - F_N$. Overall, consider (27), and given β'_{t-1} for the different distributions and β_{t-1} for the same distribution, we have $\beta'_{t-1} > \beta_{t-1}$.

Case 2. $F_B > F_R$. Denote the sum of the moving probabilities with $F_B > F_R$ by A''_{t-1} and the sum with $F_B = F_R$ by A_{t-1}. Since by Lemma 4, $A_{t-1} > 1$, we have $\alpha + (1-\alpha)\big(\mathcal{P}_{t-1}^B + \mathcal{P}_{t-1}^R\big) > 1$. Then, given $1 - F_B(k) + F_R(k) < 1$, by comparing $F_B > F_R$ with $F_B = F_R$, we have

$$\alpha[1 - F_B(k) + F_R(k)] + (1-\alpha)A''_{t-1} < \alpha + (1-\alpha)A_{t-1},$$

which implies that $A''_{t-1}(2 - A''_{t-1}) > A_{t-1}(2 - A_{t-1})$ for all t since the function $a(2-a)$ is strictly decreasing in $a > 1$. Further, denote $F''_N = F_X(k) + F_Y(k)$ for the different

distributions and $F_N = F_X(k) + F_X(k)$ for the same distribution. By FOSD between F_X and F_Y, we have $1 - F''_N < 1 - F_N$. Overall, consider (27), and given β''_{t-1} for the different distributions and β_{t-1} for the same distribution, we have $\beta''_{t-1} < \beta_{t-1}$.

The result can be shown by comparing different values of β_{t-1} in f_{t-1} from (26). We denote f_{t-1} given $F_B = F_R$ and say that a function g_{t-1} uniformly dominates f_{t-1} if $g_{t-1}(z) > f_{t-1}(z)$ for all z. The other case, $A_0 < 1$, can be shown using a similar procedure.

(i) If $A_0 > 1$ and f_{t-1} is given the symmetric distribution, then

$$\begin{cases} g_{t-1} \text{ dominates } f_{t-1} \text{ with } (F_Y, F_Y) \text{ if } g_{t-1} \text{ is given } F_B < F_R, \\ f_{t-1} \text{ dominates } g_{t-1} \text{ with } (F_X, F_X) \text{ if } g_{t-1} \text{ is given } F_B > F_R. \end{cases}$$

(ii) If $A_0 < 1$ and f_{t-1} given the symmetric distribution, then

$$\begin{cases} g_{t-1} \text{ dominates } f_{t-1} \text{ with } (F_Y, F_Y) \text{ if } g_{t-1} \text{ is given } F_B > F_R, \\ f_{t-1} \text{ dominates } g_{t-1} \text{ with } (F_X, F_X) \text{ if } g_{t-1} \text{ is given } F_B < F_R. \end{cases}$$

The dominance implies a lower fixed point. ■

References

Allen, T. and Donaldson, D. (2020), Persistence and path dependence in the spatial economy, National Bureau of Economic Research Working Paper No. 28059.

Becker, G.S. (1973), A theory of marriage: Part I, *Journal of Political Economy* 81, 813–846.

Callander, S. and Carbajal, J.C. (2022), Cause and effect in political polarization: A dynamic analysis, *Journal of Political Economy*, 130, 825–880.

Carlsson, H., and van Damme, E. (1993), Global games and equilibrium selection, *Econometrica,* 61, 989–1018.

Chen, Y. and Li, S. (2009), Group identity and social preferences. *American Economic Review* 99, 431–457.

Currarini, S., Jackson, M.O and Pin, P. (2010), Identifying the roles of race-based choice and chance in high school friendship network formation, *Proceedings of the National Academy of Sciences* 107, 4857–4861.

Damiano, E. and Li, H. (2007), Price discrimination and efficient matching, *Economic Theory* 30, 243–263.

Goeree, J.K., McConnell, M.A., Mitchell, T. Tromp, T. and Yariv, L (2010) The 1/d law of giving, *American Economic Journal: Microeconomics* 2, 183-203.

Hofbauer, J. and Sorger, G. (1999), Perfect foresight and equilibrium selection in symmetric potential games, *Journal of Economic Theory* 85, 1–23.

Huckfeldt, R. and Sprague, J. (1995), Citizens, politics and social communication: Information and influence in an election campaign, *Cambridge University Press*, 1995.

Jackson, M.O. (2014), Networks in the understanding of economic behaviors, *Journal of Economic Perspectives* 28, 3–22.

Kleinman, B., Liu, E. and Redding, S.J, (2023), Dynamic spatial general equilibrium, *Econometrica*, 91, 385–424.

Kossinets, G. and Watts, D.J. (2009), Origins of homophily in an evolving social network, *American Journal of Sociology* 115, 405–450.

Matsui, A. and Matsuyama, K. (1995), An approach to equilibrium selection, *Journal of Economic Theory* 65, 415–434.

Milgrom, P. and Shannon, C. (1994), Monotone comparative statics, *Econometrica* 62, 157–180.

McPherson, M., Smith-Lovin, L. and Cook, J.M. (2001), Birds of a feather: homophily in social networks, *Annual Review of Sociology* 27, 415–444.

Schelling, T. (1971), Dynamic models of segregation, *Journal of Mathematical Sociology* 1, 143–186.

Sherif, Muzafer, Harvey, O.J., White, B. Jack, Hood, William R., and Sherif, Carolyn W. (1961), *Intergroup conflict and cooperation: The Robbers Cave experiment* (Vol. 10). Norman, OK: University Book Exchange.

Tarski, A. (1955), A lattice theoretical fixed point theorem and its applications, *Pacific Journal of Mathematics* 5, 285–309.

Tajfel, H. (1974), Social identity and intergroup behaviour. *Social Science Information* 13, 65–93.

Tiebout, C.M. (1956), A pure theory of local expenditures, *Journal of Political Economy* 64, 416–424.

Yoo, S.H. (2014), Learning a population distribution, *Journal of Economic Dynamics & Control* 48, 188–201.

Zhuravskaya, E., Petrova, M. and Enikolopov, R. (2020), Political effects of the internet and social media, *Annual Review of Economics* 12, 415–438.

Appendix B Additional Material for Online Appendix

B.1. A finite number of locations

Denote a group g player's location strategy by $\ell_t^g : \Omega \to \{\ell_1, ..., \ell_L\}$, and further, the probability that group g members move to a location ℓ by $P_t^{g,\ell}(\omega, \ell_t^g)$ given a history ω, generalizing $P_t^g(\omega, \ell_t^g)$ in Section 3. We need only two sets of locations.

Proposition 9 *Suppose (A1)-(A4). Then, the payoff difference between a homogeneous match and a heterogeneous match is equivalent to the difference in their corresponding thresholds such that* $U_t^S(\theta, \omega^S, \mathbf{k}_t^S) - U_t^A(\theta, \omega^A, \mathbf{k}_t^A) = d(k_t^A(\omega^A)) - d(k_t^S(\omega^S))$*, and in any location equilibrium for all* $t = 1, 2, ...$*, there are only two sets of locations* E *and* W *such that for* $g \neq g' \in \{B, R\}$*, each player with* $\omega \in \Omega^{g+}$ *and* $\omega \in \Omega^{g'-}$ *moves to one location that belongs to* E*, and for each* $\ell, \widehat{\ell} \in E$*,*

$$\frac{P_t^{B,\ell}(\omega, \ell_t^B)}{P_t^{R,\ell}(\omega, \ell_t^R)} = \frac{P_t^{B,\widehat{\ell}}(\omega, \ell_t^B)}{P_t^{R,\widehat{\ell}}(\omega, \ell_t^R)},$$

and each player with $\omega \in \Omega^{g-}$ *and* $\omega \in \Omega^{g'+}$ *moves to one location that belongs to* W*, and for each* $\ell, \widehat{\ell} \in W$*,*

$$\frac{P_t^{B,\ell}(\omega, \ell_t^B)}{P_t^{R,\ell}(\omega, \ell_t^R)} = \frac{P_t^{B,\widehat{\ell}}(\omega, \ell_t^B)}{P_t^{R,\widehat{\ell}}(\omega, \ell_t^R)}.$$

Proof. If a B player chooses ℓ, he obtains the expected payoff

$$V_t^B(\ell, \theta, \omega, \ell_t) = \frac{P_t^{B,\ell}(\omega, \ell_t^B)}{P_t^{B,\ell}(\omega, \ell_t^B) + P_t^{R,\ell}(\omega, \ell_t^R)} U_t^S(\theta, \omega^S, \mathbf{k}_t^S) + \frac{P_t^{R,\ell}(\omega, \ell_t^R)}{P_t^{B,\ell}(\omega, \ell_t^B) + P_t^{R,\ell}(\omega, \ell_t^R)} U_t^A(\theta, \omega^A, \mathbf{k}_t^A),$$

where $\ell_t \equiv (\ell_t^B, \ell_t^R)$ and $U_t^S(\theta, \omega^S, \mathbf{k}_t^S)$ and $U_t^A(\theta, \omega^A, \mathbf{k}_t^A)$ are from (18) and (20) in the investment stage. If, on the other hand, the B player chooses $\widehat{\ell}$, he obtains the expected payoff

$$V_t^B(\widehat{\ell}, \theta, \omega, \ell_t) = \frac{P_t^{B,\widehat{\ell}}(\omega, \ell_t^B)}{P_t^{B,\widehat{\ell}}(\omega, \ell_t^B) + P_t^{R,\widehat{\ell}}(\omega, \ell_t^R)} U_t^S(\theta, \omega^S, \mathbf{k}_t^S) + \frac{P_t^{R,\widehat{\ell}}(\omega, \ell_t^R)}{P_t^{B,\widehat{\ell}}(\omega, \ell_t^B) + P_t^{R,\widehat{\ell}}(\omega, \ell_t^R)} U_t^A(\theta, \omega^A, \mathbf{k}_t^A).$$

The difference between group B proportion in location ℓ and that in $\widehat{\ell}$ is rewritten as

$$\begin{aligned} &\frac{P_t^{B,\ell}(\omega, \ell_t^B)}{P_t^{B,\ell}(\omega, \ell_t^B) + P_t^{R,\ell}(\omega, \ell_t^R)} - \frac{P_t^{B,\widehat{\ell}}(\omega, \ell_t^B)}{P_t^{B,\widehat{\ell}}(\omega, \ell_t^B) + P_t^{R,\widehat{\ell}}(\omega, \ell_t^R)} \\ &= \frac{P_t^{B,\ell}(\omega, \ell_t^B) P_t^{R,\widehat{\ell}}(\omega, \ell_t^R) - P_t^{B,\widehat{\ell}}(\omega, \ell_t^B) P_t^{R,\ell}(\omega, \ell_t^R)}{\left[P_t^{B,\ell}(\omega, \ell_t^B) + P_t^{R,\ell}(\omega, \ell_t^R)\right]^2}, \end{aligned}$$

and the difference between group R proportion in location ℓ and that in $\widehat{\ell}$ is

$$\begin{aligned}&\frac{P_t^{R,\ell}(\omega,\ell_t^B)}{P_t^{B,\ell}(\omega,\ell_t^B)+P_t^{R,\ell}(\omega,\ell_t^R)}-\frac{P_t^{R,\widehat{\ell}}(\omega,\ell_t^B)}{P_t^{B,\widehat{\ell}}(\omega,\ell_t^B)+P_t^{R,\widehat{\ell}}(\omega,\ell_t^R)}\\&=\frac{P_t^{B,\widehat{\ell}}(\omega,\ell_t^B)P_t^{R,\ell}(\omega,\ell_t^R)-P_t^{B,\ell}(\omega,\ell_t^B)P_t^{R,\widehat{\ell}}(\omega,\ell_t^R)}{\left[P_t^{B,\ell}(\omega,\ell_t^B)+P_t^{R,\ell}(\omega,\ell_t^R)\right]^2}.\end{aligned}$$

Denote

$$\Delta P_t^{\ell,\widehat{\ell}}(\omega,\ell_t)\equiv\frac{P_t^{B,\ell}(\omega,\ell_t^B)P_t^{R,\widehat{\ell}}(\omega,\ell_t^R)-P_t^{B,\widehat{\ell}}(\omega,\ell_t^B)P_t^{R,\ell}(\omega,\ell_t^R)}{\left[P_t^{B,\ell}(\omega,\ell_t^B)+P_t^{R,\ell}(\omega,\ell_t^R)\right]^2}$$

With $\Delta P_t^{\ell,\widehat{\ell}}(\omega,\ell_t)$, the difference in the expected payoffs can be derived as

$$\begin{aligned}V_t^B(\ell,\theta,\omega,\ell_t)-V_t^B(\widehat{\ell},\theta,\omega,\ell_t)&=\Delta P_t^{\ell,\widehat{\ell}}(\omega,\ell_t)U_t^S(\theta,\omega^S,\mathbf{k}_t^S)-\Delta P_t^{\ell,\widehat{\ell}}(\omega,\ell_t)U_t^A(\theta,\omega^A,\mathbf{k}_t^A)\\&=\Delta P_t^{\ell,\widehat{\ell}}(\omega,\ell_t)\left[U_t^S(\theta,\omega^S,\mathbf{k}_t^S)-U_t^A(\theta,\omega^A,\mathbf{k}_t^A)\right]\\&=\Delta P_t^{\ell,\widehat{\ell}}(\omega,\ell_t)\left[d(k_t^A(\omega^A))-d(k_t^S(\omega^S))\right],\end{aligned}$$

where $\Delta P_t^{\ell,\widehat{\ell}}(\omega,\ell_t)$ can be rewritten as

$$\Delta P_t^{\ell,\widehat{\ell}}(\omega,\ell_t)=\frac{P_t^{R,\ell}(\omega,\ell_t^R)P_t^{R,\widehat{\ell}}(\omega,\ell_t^R)\left[\frac{P_t^{B,\ell}(\omega,\ell_t^B)}{P_t^{R,\ell}(\omega,\ell_t^R)}-\frac{P_t^{B,\widehat{\ell}}(\omega,\ell_t^B)}{P_t^{R,\widehat{\ell}}(\omega,\ell_t^R)}\right]}{\left[P_t^{B,\ell}(\omega,\ell_t^B)+P_t^{R,\ell}(\omega,\ell_t^R)\right]^2}.$$

Hence, for ω satisfying $d(k_t^A(\omega^A))-d(k_t^S(\omega^S))\neq 0$, if group $g\in\{B,R\}$ members are located in two ℓ and $\widehat{\ell}$, the following equality must hold:

$$\frac{P_t^{B,\ell}(\omega,\ell_t^B)}{P_t^{R,\ell}(\omega,\ell_t^R)}=\frac{P_t^{B,\widehat{\ell}}(\omega,\ell_t^B)}{P_t^{R,\widehat{\ell}}(\omega,\ell_t^R)},$$

which shows the results. ■

By Proposition 9, we need to consider just two sets of locations; or simply two locations E and W. Suppose that there are five locations, 1 - 5. Then, in equilibrium, it must be the case that there are only two sets of locations satisfying the ratio conditions. For instance, location 1 and 3 have the same ratio (more Blue members) while locations 2 and 5 have another ratio that is the same (more Red members), and there is no one in location 4. Without loss of generality, we can call 1 and 3 "East" and 2 and 5 "West."

B.2. On mixed location strategies

An optimal location strategy ℓ_t is robust if it can survive a small perturbation as can be found in Section 2. Now, consider mixed location strategies for both group members $\ell_t^B:\Omega\times[-1,1]\to\Delta(L)$ and $\ell_t^R:\Omega\times[-1,1]\to\Delta(L)$ and note that a mixed location strategy equilibrium arises only when $\Delta P_t(\omega,\ell_t)=0$. Denote by $\xi^B(\omega)\in[0,1]$ for $P_t^B(\omega,\ell_t^B)$

the proportion of group B moving to E from a mixed strategy and by $\xi^R(\omega) \in [0,1]$ for $P_t^R(\omega, \ell_t^R)$ the proportion of group R moving to E from a mixed strategy. With them, $P_t^B(\omega, \ell_t^B)$ and $P_t^R(\omega, \ell_t^B)$ can be derived such that

$$\begin{aligned}
&\mathbb{E}\big[\mathbf{1}_{\{\ell_t^B(\widetilde{\omega}, z_{t-1})=E\}} \mid \omega\big] \\
&= q_{t-1}\xi^B(I,\emptyset)p_I(\omega^B)\mathbf{1}_{\{\ell_t^B((I,\emptyset),z_{t-1})=E\}} + (1-q_{t-1})\xi^B(\emptyset,N)p_N(\omega^R)\mathbf{1}_{\{\ell_t^B((\emptyset,N),z_{t-1})=E\}} \\
&\quad + q_{t-1}\xi^B(N,\emptyset)p_N(\omega^B)\mathbf{1}_{\{\ell_t^B((N,\emptyset),z_{t-1})=E\}} + (1-q_{t-1})\xi^B(\emptyset,I)p_I(\omega^R)\mathbf{1}_{\{\ell_t^B((\emptyset,I),z_{t-1})=E\}},
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\big[\mathbf{1}_{\{\ell_t^R(\widetilde{\omega}, z_{t-1})=E\}} \mid \omega\big] \\
&= (1-q_{t-1})\xi^R(I,\emptyset)p_I(\omega^B)\mathbf{1}_{\{\ell_t^R((I,\emptyset),z_{t-1})=E\}} + q_{t-1}\xi^R(\emptyset,N)p_N(\omega^R)\mathbf{1}_{\{\ell_t^R((\emptyset,N),z_{t-1})=E\}} \\
&\quad + (1-q_{t-1})\xi^R(N,\emptyset)p_N(\omega^B)\mathbf{1}_{\{\ell_t^R((N,\emptyset),z_{t-1})=E\}} + q_{t-1}\xi^R(\emptyset,I)p_I(\omega^R)\mathbf{1}_{\{\ell_t^R((\emptyset,I),z_{t-1})=E\}}.
\end{aligned}$$

Now, suppose there exists a location strategy $\ell_t = (\ell_t^B, \ell_t^R)$ with $\Delta P_t(\omega, \ell_t) = 0$ for $\Delta\mathcal{P}_{t-1} \neq 0$ since a mixed location strategy equilibrium occurs only given $\Delta P_t(\omega, \ell_t) = 0$. By taking the difference between $P_t^B(\omega, \ell_t^B)$ and $P_t^R(\omega, \ell_t^B)$, we have

$$0 = \Delta P_t(\omega, \ell_t) = \alpha\left\{\mathbb{E}\big[\mathbf{1}_{\{\ell_t^B(\widetilde{\omega}, z_{t-1})=E\}} \mid \omega\big] - \mathbb{E}\big[\mathbf{1}_{\{\ell_t^B(\widetilde{\omega}, z_{t-1})=E\}} \mid \omega\big]\right\} + (1-\alpha)\Delta\mathcal{P}_{t-1}.$$

This means that for some $\Delta\mathcal{P}_{t-1} \neq 0$, there exist $\ell_t^B : \Omega \times [-1,1] \to \Delta(L)$ and $\ell_t^R : \Omega \times [-1,1] \to \Delta(L)$ satisfying the equality above. Note that this holds for a particular α value, so it is not robust. This implies that for a robust location strategy equilibrium, $\Delta P_t(\omega, \ell_t) = 0$ if and only if $\Delta\mathcal{P}_{t-1} = 0$. Hence, for $\Delta\mathcal{P}_{t-1} \neq 0$, we have $\Delta P_t(\omega, \ell_t) \neq 0$ for which there is no mixed location strategy following (22).

B.3. The proof of Lemma 5

Proof of Lemma 5. *(i), (ii)* Take the difference between $k_t^A(I|\emptyset)$ and $k_t^A(\emptyset|I)$ such that

$$\begin{aligned}
&d(k_t^A(I|\emptyset)) - d(k_t^A(\emptyset|I)) \\
&= \pi(I^R)X_t^A(\mathbf{k}_t^A, \emptyset^B) + (1-\pi(I^R))Y_t^A(\mathbf{k}_t^A, \emptyset^B) - [\pi(\emptyset^R)X_t^A(\mathbf{k}_t^A, I^B) + (1-\pi(\emptyset^R))Y_t^A(\mathbf{k}_t^A, I^B)],
\end{aligned}$$

where the RHS can be expanded as follows:

$$\pi(I^R)\begin{bmatrix} q_{t-1}(1-F_X(k))F_X(k_t^A(\emptyset|I)) + q_{t-1}F_X(k)F_X(k_t^A(\emptyset|N)) \\ +(1-q_{t-1})\big[\pi(\emptyset^B)(1-F_X(k)) + (1-\pi(\emptyset^B))(1-F_Y(k))\big]F_X(k_t^A(I|\emptyset)) \\ +(1-q_{t-1})\big[\pi(\emptyset^B)F_X(k) + (1-\pi(\emptyset^B))F_Y(k)\big]F_X(k_t^A(N|\emptyset)) \end{bmatrix} \tag{49}$$

$$+(1-\pi(I^R))\begin{bmatrix} q_{t-1}(1-F_Y(k))F_Y(k_t^A(\emptyset|I)) + q_{t-1}F_Y(k)F_Y(k_t^A(\emptyset|N)) \\ +(1-q_{t-1})\big[\pi(\emptyset^B)(1-F_X(k)) + (1-\pi(\emptyset^B))(1-F_Y(k))\big]F_Y(k_t^A(I|\emptyset)) \\ +(1-q_{t-1})\big[\pi(\emptyset^B)F_X(k) + (1-\pi(\emptyset^B))F_Y(k)\big]F_Y(k_t^A(N|\emptyset)) \end{bmatrix}$$

$$-\pi(\emptyset^R)\begin{bmatrix} q_{t-1}(1-F_X(k))F_X(k_t^A(\emptyset|I)) + q_{t-1}F_X(k)F_X(k_t^A(\emptyset|N)) \\ +(1-q_{t-1})\big[\pi(I^B)(1-F_X(k)) + (1-\pi(I^B))(1-F_Y(k))\big]F_X(k_t^A(I|\emptyset)) \\ +(1-q_{t-1})\big[\pi(I^B)F_X(k) + (1-\pi(I^B))F_Y(k)\big]F_X(k_t^A(N|\emptyset)) \end{bmatrix} \tag{50}$$

$$-(1-\pi(\emptyset^R))\begin{bmatrix} q_{t-1}(1-F_Y(k))F_Y(k_t^A(\emptyset|I)) + q_{t-1}F_Y(k)F_Y(k_t^A(\emptyset|N)) \\ +(1-q_{t-1})\big[\pi(I^B)(1-F_X(k)) + (1-\pi(I^B))(1-F_Y(k))\big]F_Y(k_t^A(I|\emptyset)) \\ +(1-q_{t-1})\big[\pi(I^B)F_X(k) + (1-\pi(I^B))F_Y(k)\big]F_Y(k_t^A(N|\emptyset)) \end{bmatrix}.$$

Since $\pi(I^R)\pi(\emptyset^B) = \pi(I^B)\pi(\emptyset^R)$, the third term of the bracket in (49) and the third term of the bracket in (50) can be simplified such that

$$\begin{aligned} &\pi(I^R)(1-q_{t-1})\big[\pi(\emptyset^B)(1-F_X(k)) + (1-\pi(\emptyset^B))(1-F_Y(k))\big]F_X(k_t^A(I|\emptyset)) \\ &-\pi(\emptyset^R)(1-q_{t-1})\big[\pi(I^B)(1-F_X(k)) + (1-\pi(I^B))(1-F_Y(k))\big]F_X(k_t^A(I|\emptyset)) \\ &=\pi(I^R)(1-q_{t-1})(1-F_Y(k))F_X(k_t^A(I|\emptyset)) - \pi(\emptyset^R)(1-q_{t-1})(1-F_Y(k))F_X(k_t^A(I|\emptyset)). \end{aligned} \tag{51}$$

The same type of deletion can be applied to the fourth term of the bracket in (49) and the fourth term of the bracket in (50). Furthermore, considering the bracket following $(1-\pi(I^R))$ and the bracket following $(1-\pi(\emptyset^R))$, the whole formula can be rewritten as

$$\begin{aligned} &\pi(I^R)\begin{bmatrix} q_{t-1}(1-F_X(k))F_X(k_t^A(\emptyset|I)) + q_{t-1}F_X(k)F_X(k_t^A(\emptyset|N)) \\ +(1-q_{t-1})(1-F_Y(k))F_X(k_t^A(I|\emptyset)) + (1-q_{t-1})F_Y(k)F_X(k_t^A(N|\emptyset)) \end{bmatrix} \\ &+(1-\pi(I^R))[q_{t-1}(1-F_Y(k))F_Y(k_t^A(\emptyset|I)) + q_{t-1}F_Y(k)F_Y(k_t^A(\emptyset|N))] \\ &-\pi(\emptyset^R)\begin{bmatrix} q_{t-1}(1-F_X(k))F_X(k_t^A(\emptyset|I)) + q_{t-1}F_X(k)F_X(k_t^A(\emptyset|N)) \\ +(1-q_{t-1})(1-F_Y(k))F_X(k_t^A(I|\emptyset)) + (1-q_{t-1})F_Y(k)F_X(k_t^A(N|\emptyset)) \end{bmatrix} \\ &-(1-\pi(\emptyset^R))[q_{t-1}(1-F_Y(k))F_Y(k_t^A(\emptyset|I)) + q_{t-1}F_Y(k)F_Y(k_t^A(\emptyset|N))] \\ &+(1-q_{t-1})(\pi(\emptyset^R) - (\pi(I^R))(1-F_X(k))F_Y(k_t^A(I|\emptyset)) \\ &+(1-q_{t-1})(\pi(\emptyset^R) - (\pi(I^R))F_X(k)F_Y(k_t^A(N|\emptyset)), \end{aligned}$$

where the last two terms are rewritten from terms with $(1-q_{t-1})$ in the bracket following $(1-\pi(I^R))$ and in the bracket following $(1-\pi(\emptyset^R))$, using $(1-\pi(I^R))(1-\pi(\emptyset^B)) = (1-\pi(I^B))(1-\pi(\emptyset^R))$ and a procedure similar to (51). Additionally, we simplify terms for $F_Y(k_t^A(I|\emptyset))$ and those for $F_Y(k_t^A(N|\emptyset))$ such that

$$\begin{aligned} &(1-\pi(I^R))(1-q_{t-1})\big[\pi(\emptyset^B)(1-F_X(k)) + (1-\pi(\emptyset^B))(1-F_Y(k))\big]F_Y(k_t^A(I|\emptyset)) \\ &-(1-\pi(\emptyset^R))(1-q_{t-1})\big[\pi(I^B)(1-F_X(k)) + (1-\pi(I^B))(1-F_Y(k))\big]F_Y(k_t^A(I|\emptyset)) \\ &=(1-q_{t-1})(\pi(\emptyset^R) - (\pi(I^R))(1-F_X(k))F_Y(k_t^A(I|\emptyset)), \end{aligned}$$

and also

$$\begin{aligned}
&(1-\pi(I^R))(1-q_{t-1})\big[\pi(\emptyset^B)F_X(k)+(1-\pi(\emptyset^B))F_Y(k)\big]F_Y(k_t^A(N|\emptyset))\\
&-(1-\pi(\emptyset^R))(1-q_{t-1})\big[\pi(I^B)F_X(k)+(1-\pi(I^B))F_Y(k)\big]F_Y(k_t^A(N|\emptyset))\\
&=(1-q_{t-1})(\pi(\emptyset^R)-(\pi(I^R))F_X(k)F_Y(k_t^A(N|\emptyset)).
\end{aligned}$$

Then, the above is rewritten as

$$\begin{aligned}
&\pi(I^R)X_t^A(\mathbf{k}_t^A,\emptyset^B)+(1-\pi(I^R))Y_t^A(\mathbf{k}_t^A,\emptyset^B)-[\pi(\emptyset^R)X_t^A(\mathbf{k}_t^A,I^B)+(1-\pi(\emptyset^R))Y_t^A(\mathbf{k}_t^A,I^B)]\\
&=\pi(I^R)\left[\begin{array}{l} q_{t-1}(1-F_X(k))F_X(k_t^A(\emptyset|I))+q_{t-1}F_X(k)F_X(k_t^A(\emptyset|N))\\ +(1-q_{t-1})(1-F_Y(k))F_X(k_t^A(I|\emptyset))+(1-q_{t-1})F_Y(k)F_X(k_t^A(N|\emptyset))\end{array}\right]\\
&-\pi(\emptyset^R)\left[\begin{array}{l} q_{t-1}(1-F_X(k))F_X(k_t^A(\emptyset|I))+q_{t-1}F_X(k)F_X(k_t^A(\emptyset|N))\\ +(1-q_{t-1})(1-F_Y(k))F_X(k_t^A(I|\emptyset))+(1-q_{t-1})F_Y(k)F_X(k_t^A(N|\emptyset))\end{array}\right]\\
&-q_{t-1}(\pi(I^R)-\pi(\emptyset^R))[(1-F_Y(k))F_Y(k_t^A(\emptyset|I))+F_Y(k)F_Y(k_t^A(\emptyset|N))]\\
&-(1-q_{t-1})(\pi(I^R)-(\pi(\emptyset^R))[(1-F_X(k))F_Y(k_t^A(I|\emptyset))+F_X(k)F_Y(k_t^A(N|\emptyset))],
\end{aligned}$$

which is

$$\begin{aligned}
&q_{t-1}(\pi(I^R)-\pi(\emptyset^R))[(1-F_X(k))F_X(k_t^A(\emptyset|I))+F_X(k)F_X(k_t^A(\emptyset|N))]\\
&+(1-q_{t-1})(\pi(I^R)-(\pi(\emptyset^R))[(1-F_Y(k))F_X(k_t^A(I|\emptyset))+F_Y(k)F_X(k_t^A(N|\emptyset))]\\
&-q_{t-1}(\pi(I^R)-\pi(\emptyset^R))[(1-F_Y(k))F_Y(k_t^A(\emptyset|I))+F_Y(k)F_Y(k_t^A(\emptyset|N))]\\
&-(1-q_{t-1})(\pi(I^R)-(\pi(\emptyset^R))[(1-F_X(k))F_Y(k_t^A(I|\emptyset))+F_X(k)F_Y(k_t^A(N|\emptyset))]\\
&<q_{t-1}(\pi(I^R)-\pi(\emptyset^R))(F_Y(k)-F_X(k))\left[F_r(k_t^A(\emptyset|I))-F_r(k_t^A(\emptyset|N))\right]\text{ for all } r\in\{X,Y\},
\end{aligned}$$

where the inequality follows from the FOSD between F_X and F_Y. This establishes (i). Similarly, (ii) can be readily shown.

(iii) Now, take the difference between $k_t^A(\emptyset|I)$ and $k_t^A(\emptyset|N)$ such that

$$\begin{aligned}
&d(k_t^A(\emptyset|I))-d(k_t^A(\emptyset|N))\\
&=\pi(\emptyset^R)X_t^A(\mathbf{k}_t^A,I^B)+(1-\pi(\emptyset^R))Y_t^A(\mathbf{k}_t^A,I^B)-[\pi(\emptyset^R)X_t^A(\mathbf{k}_t^A,N^B)+(1-\pi(\emptyset^R))Y_t^A(\mathbf{k}_t^A,N^B)]\\
&=\pi(\emptyset^R)(1-q_{t-1})\left[F_X(k_t^A(I|\emptyset))-F_X(k_t^A(N|\emptyset))\right]\big(\pi(I^B)-\pi(N^B)\big)[F_Y(k)-F_X(k)]\\
&\quad+(1-\pi(\emptyset^R))(1-q_{t-1})\left[F_Y(k_t^A(I|\emptyset))-F_Y(k_t^A(N|\emptyset))\right]\big(\pi(I^B)-\pi(N^B)\big)[F_Y(k)-F_X(k)]\\
&=\frac{1}{2}(1-q_{t-1})\big(\pi(I^B)-\pi(N^B)\big)[F_Y(k)-F_X(k)]\left[\begin{array}{l} F_X(k_t^A(I|\emptyset))-F_X(k_t^A(N|\emptyset))\\ +F_Y(k_t^A(I|\emptyset))-F_Y(k_t^A(N|\emptyset))\end{array}\right],
\end{aligned}$$

which establishes (iii). ■