On the Origin of Polarization*

John Duffy† Seung Han Yoo‡

July 2022

Abstract

We provide a model of group sorting or polarization based on group identity alone. In our model, agents differ from one another in terms of an observable binary group identity. Groups may also differ in terms of the distribution of abilities (types) but the true distribution is uncertain, so agents have to form beliefs about that distribution in making both investment and location decisions. Each agent's ability is private information, whereas group identity is publicly observable. Agents have no preferences or special facilities for interacting with members of their own group. In this environment, we show that, in equilibrium, agents endogenously sort themselves according to their group identity to two different locations under rational belief updating, and we identify conditions under which 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 Joel Sobel and Leeat Yariv for very helpful comments. The usual disclaimer applies.

†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

People like to interact with those who are similar to themselves. Such preferences for *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] However, there is no theory that explains how and why such sorting arises that is not preference-based but is instead based on one group's *perception* of the other group. In this paper, we develop such an alternative theory showing how perfect sorting of group members to different locations, or polarization can be perception-based. One can think of our model as providing a kind of "statistical discrimination" rationalization for polarization as opposed to the "taste-based discrimination" approach to polarization as 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 belong to one of two groups, labeled Red and Blue. 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. 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, under rational belief updating.[2]

As with all origin stories, we need a plot device that does not strain credulity. The mechanism we employ is distributional uncertainty together with private monitoring and Bayesian belief updating. Specifically, we consider the case where there are two possible distributions for each group's types 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 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. Both young and old players participate in an investment game with members of their own generation. Young agents play the game in the location chosen by their parent and then decide whether to remain in the 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 players in the two locations at each period affect the *matching probabilities*, which

[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, McConnell, Mitchell and Yariv (2010).

[2]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.

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.

We start by providing a simple and yet powerful first result, which we term the natural law of likes meeting likes: In any generic, non-strategic setting, it is more likely for a player to meet a member of his or her own group than a member of the other group. We then characterize the investment stage game outcomes from homogeneous and heterogeneous matches between players followed by a characterization of the location stage equilibrium. We find sets of histories that are favorable and unfavorable for playing the game with members of one's own group, which together with the law of likes meeting likes generates an endogenous location choice dynamical system. This system consists of a simple convex (square) function characterizing the difference in the location choice probabilities of the two groups. The square function nicely captures the vulnerability and risk that the society becomes polarized.

To illustrate, suppose that players in the location choice stage of the game 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 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 different group member, a Blue member. The former has the effect of widening the polarization of the two groups, whereas the latter has the effect of reducing this gap. Yet, the former effect dominates the latter due to the natural law of likes meeting likes – the likelihood of a meeting between two Blue members in youth is higher than that of a meeting between a Blue and Red member. Consider next the history where a player observes No Invest from a match with an Red player when young, a "bad" outcome. In fact, this history works the same as in the previous example since the No Invest by a Red member is a relatively favorable history towards a player's own group if the player matched to the Red member is a Blue member, but it is an unfavorable history toward the player's own group if the player matched to the Red member is also an Red member, and so both players will choose East. This time, however, the polarization-reducing effect dominates the polarization-amplifying effect due to the same law of likes meeting likes since a matching between two Red members in youth is more frequent than a Red-Blue match.

Hence, if the overall probability of generating the Invest equilibrium is greater than the probability of generating the No Invest equilibrium, the location equilibrium yields a dynamical system in which polarization is increasing. Yet, the convergence outcome is *subtle*: the dynamical system may not result in polarization, despite the strict monotonicity. Specifically, without any exogenous force in each period, the state variable representing the difference in location choice probabilities is always smaller than the period-specific *fixed point of the system characterizing the population difference*, which is always moving over time. In the absence of any frictions, the limiting outcome is a completely mixed state where both Red and Blue group members can be found in one or both

locations.

However, suppose that only some proportion of agents make endogenous location decisions; the remaining proportion follow exogenous, systematic polarization forces in their parent's locations. In that case, for a given amount of systematic polarization, there exists a corresponding initial population difference between the two locations such that each period's state variable is always greater than the period's moving fixed point in every period.[3] 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 in the other location as opposed to the completely mixed state.

Our paper is related to several different literatures. First, our paper is related to the matching literature where the seminal work of Becker (1973) examines conditions under which an assortative matching arises as an equilibrium (see further contributions by Shimer and Smith (2000), Legros and Newman (2002, 2007) among others). 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. We also 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. The most closely related matching papers to this paper thus can be divided into two branches. The first strand considers a *centralized* matching setup with two-sided incomplete information. In Damiano and Li (2007), 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 achieve the second-best solution. Board (2009) and Hoppe, Moldovanu and Ozdenoren (2011) also study centralized matching with two-sided incomplete information. The approach is more suitable for two-sided markets with a principal such as a platform, but for the environment we study exploring how biased beliefs and polarization arise dynamically, a decentralized matching process is imperative. The second strand of the related matching literature combines decentralized two-sided matching models with search frictions and *complete information* about subject types resulting in the sorting of players by type, *e.g.*, Morgan (1995), Burdett and Coles (1997), and Smith (2006). In these papers, there exist matching equilibria where players form clusters and interact only with members of those clusters. As type quality is complementary in production and players are impatient, segregation improves market efficiency by reducing search costs and the negative externalities from matchings with low types. Our approach differs from this literature in that players face uncertainty about player types and engage in multiple trades over time with different partners. Players can only condition on group membership and their own histories of interactions so that belief updating plays an important role. Importantly these beliefs are are two-sided and determine where agents decide to locate. Those location choices in turn, affect the probabilities with which agents meet other agents from the two groups.

Second, our paper is related to the literature on private monitoring in dynamic, non-cooperative games (see, *e.g.*, Kandori (2002) for an introduction). Research on the learning of a population distribution under private monitoring can be found in Yoo (2014), but the analysis in that paper applies to a setting with a single group distribution and matching within that single group.

Third, our paper makes use of and advances the monotone comparative statics analysis of

[3]We restrict attention to a *Markov location equilibrium*, and also provide conditions that rationalize it.

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, which is necessary to identify a set of histories favorable toward a player's own group and 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 parameter 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

Finally, the topic of polarization has been studied by many other authors. As noted earlier, this literature mainly considers *preference-based* explanations for sorting following the seminal work of Schelling (1971) and Tiebout (1956)). For instance, the literature on "echo chambers" (see Levy and Razin (2019) for a recent survey) provides theoretical models and evidence for segregation of individuals with like-minded individuals and how such segregation impacts agents' beliefs about the merits of such segregation resulting in self-fulfilling echo chambers.

There is also a literature on network-based explanations for sorting and polarization (see Jackson (2014) for a recent survey). This literature has followed two different approaches. First, there can be complementarities in agents' beliefs as in Peski (2008) or preferences for similar strategic choices. For instance, Baccara and Yariv (2016) show in an endogenous network formation game setting, how polarized groups consisting of extreme partisans for either of two public projects can mitigate free-riding problems. Second, agents can have abilities to communicate or coordinate with other agents of the same type, *e.g.*, Galeotti, Ghiglino and Squintani (2013), Calvó-Armengol, De Martí and Prat (2015). For instance, Kets and Sandroni (2019) suppose that a player's own mental state is more aligned with the mental states of own group members than with members of other groups so that a desire to reduce strategic uncertainty can lead to full segregation by group identity. 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 together with private monitoring and Bayesian updating that results in the sorting of players by type into two different locations.

Regarding systematic polarization, it is well documented that social media can play a role in fostering 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. Thus, 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.

We note that our approach is related to a research agenda in Sociology and Social Psychology that has sought to find *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 stereotypes, as we do here as well, so as 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.[4] 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 and we provide conditions under which either outcome can arise. The focus of our paper is, however, to identify the underlying forces leading to such outcomes, that is, the origin of polarization.

In the next section, we introduce our model. In Section 3, we discuss exogenous location choice. 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 social dynamics of our game and establish the main polarization results. Finally, Section 7 provides discussion and Section 8 concludes. All the proofs are collected in an appendix

2 Model

There are two groups, Blue and Red, denoted by $g \in \{B, R\}$, with the same group size, unit mass. Members from the two groups meet a partner to play a stage game in one of two locations, East or West, denoted by $\ell \in \{E, W\}$. All members live two periods, youth and old age and make three lifetime decisions. Both young and old make an investment decision in each period. Then, only the young members make a location decision as to where they will reside 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:[5]

> Period $t-1$: young members are born in their parent's location at time $t-1$ and are members of the *same* group g as their parent. These young members are then randomly paired with members of their same (young) generation to play a stage game in the location of their birth.
>
> Interim period: At the end of period $t-1$, young members choose whether to move to location E or W, where they will reside in old age, period t.
>
> Period t: all old members of the $t-1$ generation are randomly paired with other members of their same (old) generation to play the stage game one final time in the location they chose for old age.

[4]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.

[5]Here, the young player's parent does not necessarily mean the parent through any blood ties *per se* in a physical location. Rather, it can mean any older generation player of the same group identity who has strong (parental-like) influence over the young even in an online space.

	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

Hence, in each period $t = 1, 2, ...,$ young members are born into the same group identity and location as their parent, and the young and the old are randomly matched with members of their own generation to play the same stage game.[6] The stage game is called the *investment stage*, and the interim phase is referred to as the *location stage*.

In the investment stage, each player chooses an action a of whether to invest I or not N, which yields player i payoff $u(a_i, a_j, \theta_i)$, where θ_i (θ_j) is row (column) player i's (j's) ability or *type*. This 2×2 game, without loss of generality, can be normalized as shown in Table 1 if $u(a_i, a_j, \theta_i)$ satisfies *increasing difference*.[7] The increasing difference relationship implies that there is a strategic complementarity between two players. However, in contrast to preference-based polarization papers, the complementarity of this model satisfies *color-blindness*; that is, regardless of the matching with a same group member or with a different group member, the degree of complementarity is *identical*. We assume that $d(\cdot)$ is continuously strictly increasing in order to induce a threshold equilibrium for the stage game.

A player's type, θ_i, is private information, while his 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 $\Theta \equiv [\overline{\theta}, \underline{\theta}]$, but, differently, no player knows the true distribution in this model. Hence, there is uncertainty about each group's distribution *as well as* uncertainty about a matched player j's type. Each group's distribution can be either F_X or F_Y, so that there are four possible combinations, $\{F_X, F_Y\} \times \{F_X, F_Y\}$, for the two-group distributions (F_B, F_R). Given the observable group identity, for each meeting in location ℓ, three types of matchings are possible: two types of homogeneous (same group) matches, (B, B), (R, R), denoted by $m = S$, and one type of heterogeneous (or asymmetric group) match, (B, R), denoted by $m = A$.

The young player's payoff does not depend on any history since the young player's beliefs are not inherited from the player's parent, an old player, while his location when young is inherited from the parent. That is, young players have an unbiased belief that F_X and F_Y are equally likely for both groups or that $\pi_1 = \frac{1}{2}$, where π_1 denotes the prior belief in youth that a group's distribution is F_X.[8] A young agent i in the investment stage game anticipates that his matched

[6] By focusing on intra-generational meetings – between young and between old members – we can utilize a symmetric equilibrium, even for matching between Blue and Red, *i,e,*, heterogeneous matches, which makes our analysis tractable.

[7] 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. $d_N(\theta_i)$) denotes the loss that player i incurs when player i unilaterally deviates from the action profile (I, I) (resp. (N, N)). Then, a 2×2 game can be normalized as the stage game in the figure by $d(\theta_i) \equiv d_I(\theta_i)/(d_I(\theta_i) + d_N(\theta_i))$ 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 increasing difference, $u(I, I, \theta_i) - u(I, N, \theta_i) > u(N, I, \theta_i) - u(N, N, \theta_i)$.

[8] Henceforth, we will adopt subscripts 1 and 2 to indicate difference in beliefs, actions and histories between young and old agents.

player j chooses No Invest with probability $\Pr(N|k_j)$, where k_j is player j's threshold for investing such that player j chooses Invest if $\theta_j > k_j$; No Invest otherwise. Thus, a choice of Invest yields player i the expected payoff

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

If player i, on the other hand, chooses No Invest, his payoff is 0. With no initial bias, in each location ℓ and matching m, $\Pr(N|k_j) = \pi_1 F_X(k_j) + (1-\pi_1)F_Y(k_j)$ with $\pi_1 = \frac{1}{2}$.

The history of play from this stage game among the young not only affect each player's investment decision when he or she is old; in anticipation of a future payoff, this history of play also affects each young player's location decision. We let $\pi_2(I)$ and $\pi_2(N)$ denote an old player's belief about the probability that the distribution of a matched partner's group is F_X given that he observed I and N from the group member, respectively. By Bayes rule, an old player's beliefs are updated as follows:

$$\pi_2(I) = \frac{(1-F_X(k_1))\pi_1}{(1-F_X(k_1))\pi_1 + (1-F_Y(k_1))(1-\pi_1)}; \text{ and } \pi_2(N) = \frac{F_X(k_1)\pi_1}{F_X(k_1)\pi_1 + F_Y(k_1)(1-\pi_1)}, \tag{2}$$

whereas no match with a group yields no update for that group. Then, an old player's payoff has a payoff structure similar to (1), but the probability that his matched player j chooses No Invest depends on his or her history of play from the investment stage game when young and the belief updating in (2), which is the subject of intensive study in Section 4.

In what follows, we rely on the following four assumptions.[9]

(A1) For each $r \in \{X, Y\}$, if F_r is the true distribution, then there exists an interior k_r.
(A2) There exists a subinterval Γ of Θ such that $\Gamma \equiv \{\theta \in \Theta : F_X(\theta) < F_Y(\theta)\}$.
(A3) For each pair $\theta' > \theta$ in Γ, $d(\theta') - d(\theta) \geq \frac{1}{2}[F_X(\theta') - F_X(\theta)] + \frac{1}{2}[F_Y(\theta') - F_Y(\theta)]$.
(A4) Monitoring is private.

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 we use the first-order stochastic dominance (A2) in the local sense to define F_X as the "better" distribution on a subinterval of the support. We make assumption (A3) for two reasons that concern the young player's equilibrium outcomes, which turns out to be useful for the old player's equilibrium outcomes as well. First, by (A3), the stochatically dominant distribution, F_X, yields a *lower* threshold, $k_X < k_Y$, meaning that there is a higher probability of choosing Invest for $r = X$.[10] We may call $[k_X, k_Y]$ the *effective support* in the sense that all equilibrium thresholds arise in that interval. Second, by (A1), there exists an interior equilibrium threshold, k_1, for the young such that

$$d(k_1) = \frac{1}{2}F_X(k_1) + \frac{1}{2}F_Y(k_1). \tag{3}$$

[9]For each $r \in \{X, Y\}$, if F_r was 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 j does not invest. By (A1), there exists an interior equilibrium threshold, k_r, satisfying $d(k_r) = F_r(k_r)$ if F_r was known to be the true distribution.

[10]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 $r \in \{X, Y\}$. 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 with (A3).

Then, by (A3), there exists a *unique* k_1, and further, with the monotone mapping $d^{-1}\big(\frac{1}{2}F_X(k_1) + \frac{1}{2}F_Y(k_1)\big)$, for a given comparative statics change from (X, Y) to (X', Y') such that $F_{X'}(\theta) + F_{Y'}(\theta) < F_X(\theta) + F_Y(\theta)$, k_1 *decreases* from such a change, which is indeed 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).

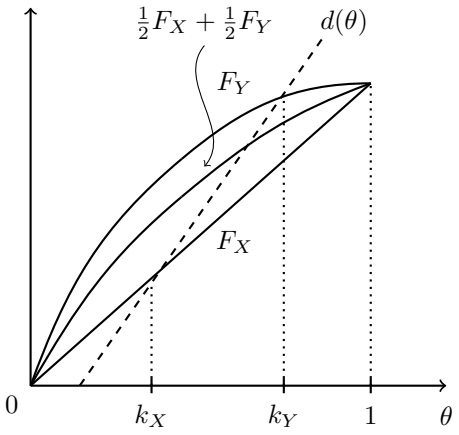

Figure 1: $d(\theta)$ and F_k and satisfying (A1)-(A3)

The private monitoring assumption (A4) means that each player observes only his own history, anticipating his partner's history. It follows that a group g old player's strategy for each $m \in \{S, A\}$ is a function of both his type θ and history ω_2^m such that $s_2(\theta, \omega_2^m) = I$ if $\theta > k_t(\omega_2^m)$ and N otherwise, where $k_t(\omega_2^m)$ denotes the old player's threshold value that depends on history ω_2^m. Note that this threshold comes with a time t subscript since the old player's equilibrium, unlike the young player's equilibrium threshold k_1, can depend on the same group *matching probability*, denoted by q_{t-1} and defined below. This matching probability serves as the *long-term memory* of the system, though players in the model have only 1 period histories and don't inherit any beliefs from their parents.

Assumption (A4) together with the type uncertainty and the likelihood of meeting a group member in either location is a distinct and important feature of our model.

3 Benchmark: exogenous location choice

Before embarking on endogenizing the probabilities with which players move to either location, we consider a benchmark case where the probability of moving to a location is *exogenously* given. We shall later use this benchmark case to derive players' location decisions *endogenously* as well as the properties of the population dynamics.

Let $\mathcal{P}_t^B$ ($\mathcal{P}_t^R$) denote the probability that B (R) group members move to meeting 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 $\frac{1-\mathcal{P}_t^B}{2-\mathcal{P}_t^B-\mathcal{P}_t^R}$. With these definitions, the overall probability, in both locations, that a

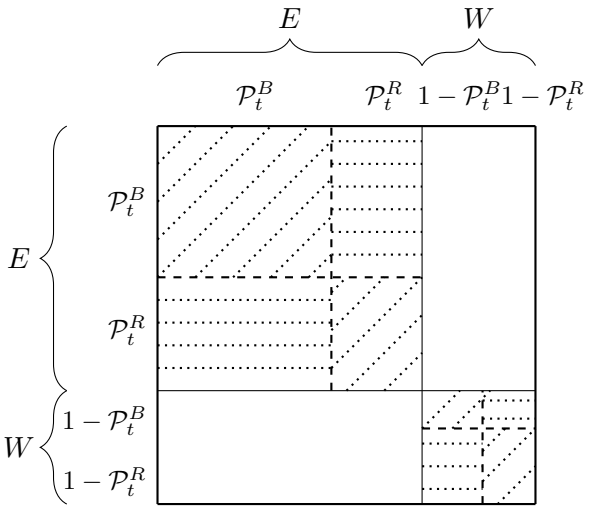

Figure 2: Same & different group matching probabilities

player is matched with a member of their *same* group, denoted by q_t, is given by

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

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

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 is powerful; yet can be straightforwardly understood. 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 a 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 $q_t^{BR} = \frac{\mathcal{P}_t^B\mathcal{P}_t^R}{\mathcal{P}_t^B+\mathcal{P}_t^R} + \frac{(1-\mathcal{P}_t^B)(1-\mathcal{P}_t^R)}{2-\mathcal{P}_t^B-\mathcal{P}_t^R}$. By the simple algebra, we have $q_t^B + q_t^R \geq (>)\, 2q_t^{BR} = 2q_t \geq (>)\, 2q_t^{BR}$ for all $\mathcal{P}_t^B$, $\mathcal{P}_t^R$ ($\mathcal{P}_t^B \neq \mathcal{P}_t^R$), as 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 $2q_t^{BR}$. In other words, there is an underlying force that generically makes matches with members of the same group more frequent. The probability q_t is a *state variable* in the subsequent development of the endogenous location decision and population dynamics of the game we study.

4 Investment stage equilibrium

The equilibrium investment choice of the young player has already been characterized by the equilibrium threshold (3) in Section 2. In this section, we focus on the old player's investment stage equilibrium behavior which requires belief updating unlike the young's. Since we search for a symmetric equilibrium for two matched players in both a homogeneous match and a heterogeneous match throughout our analysis, to simplify notations, we drop i and j from the subscripts in what follows.

4.1 Homogeneous match

We first consider the case where an old player is matched with a member of his same group, $m = S$. Similar to the young player's payoff in (1), the essential element of the old player's expected payoff from Invest is his belief about the probability that his matched partner chooses No Invest. For the old player, this probability $\Pr(N|\omega_2^S, \mathbf{k}_t^S)$ may depend on how good is the distribution of the matched partner's group – the same group in this case – and the partner's strategy of choosing No Invest given the history from youth. Thus, the group evaluation is conditional on the player's past observation. We let ω_2^S denote an old player's history for a homogeneous match, where $\omega_2^S \in \Omega_2^S \equiv \{I, \emptyset, N\}$; specifically, from the same group member, when young, if the player observed Invest, $\omega_2^S = I$, while if the player observed No Invest, $\omega_2^S = N$, and if the player was matched with a different group member, then nothing is observed from the same group, *i.e.*, $\omega_2^S = \emptyset$.

The history-contingent stage game threshold is denoted by $k_t^S(\omega_2^S)$ for $\omega_2^S \in \Omega_2^S$. In addition, we denote a homogeneous match equilibrium profile of the old player's stage game thresholds by $k_t^S \equiv \left(k_t^S(\omega_2^S)\right)_{\{\omega_2^S \in \Omega_2^S\}}$ and a profile including the youthful k_1 threshold, by $\mathbf{k}_t^S \equiv (k_1, k_t^S)$. Then, an old player in match $m = S$ obtains an expected payoff from Invest equal to

$$U_t^S(\theta, \omega_2^S, \mathbf{k}_t^S) = d(\theta) - \Pr(N|\omega_2^S, \mathbf{k}_t^S), \tag{5}$$

and, as in the young player case, No Invest yields a payoff 0. The probability of a matched partner choosing No Invest is derived as $\Pr(N|\omega_2^S, \mathbf{k}_t^S) = \pi_2(\omega_2^S)X_t^S(\mathbf{k}_t^S) + (1 - \pi_2(\omega_2^S))Y_t^S(\mathbf{k}_t^S)$, where the history-contingent probabilities are given by

$$\begin{aligned} X_t^S(\mathbf{k}_t^S) &= q_{t-1}(1 - F_X(k_1))F_X(k_t^S(I)) + q_{t-1}F_X(k_1)F_X(k_t^S(N)) + (1 - q_{t-1})F_X(k_t^S(\emptyset)), \\ Y_t^S(\mathbf{k}_t^S) &= q_{t-1}(1 - F_Y(k_1))F_Y(k_t^S(I)) + q_{t-1}F_Y(k_1)F_Y(k_t^S(N)) + (1 - q_{t-1})F_Y(k_t^S(\emptyset)). \end{aligned} \tag{6}$$

For example, given a homogeneous match (B, B), conditional on $r \in \{X, Y\}$, with probability q_{t-1} from (4), the matched B player met another B player previously when young, and with probability $(1-F_r(k_1))$, he observed Invest from that partner, where k_1 is the equilibrium stage game threshold when young from (3).[11] For the observation, a corresponding threshold is given as $k_t^S(I)$. Together, this yields the first term in (6), and the other two terms can be derived accordingly.

A homogeneous match equilibrium is defined such that for each $\ell \in \{E, W\}$ and every $\omega_2^S \in \{I, \emptyset, N\}$,

$$d(k_t^S(\omega_2^S)) = \pi_2(\omega_2^S)X_t^S(\mathbf{k}_t^S) + (1 - \pi_2(\omega_2^S))Y_t^S(\mathbf{k}_t^S). \tag{7}$$

Unlike the equilibrium for the young in (3), the old player's equilibrium is a fixed point of a *multivariable* mapping. Formally, denote $\Phi_t^S(k_t^S, \omega_2^S) \equiv d^{-1}\big(\pi_2(\omega_2^S)X_t^S(\mathbf{k}_t^S) + (1-\pi_2(\omega_2^S))Y_t^S(\mathbf{k}_t^S)\big)$, where recall that $\mathbf{k}_t^S \equiv (k_1, k_t^S)$. 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

$$\mathbf{\Phi}_t^S(k_t^S) \equiv \left(\Phi_t^S(k_t^S, \omega_2^S)\right)_{\{\omega_2^S \in \Omega_2^S\}}. \tag{8}$$

[11] The probability of observing any history is not location specific, precisely because all the young members choose the same k_1. For instance, $q_{t-1}(1 - F_X(k_1))$ in the first term of $X_t^S(\mathbf{k}_t^S)$ is given by

$$\mathcal{P}_{t-1}^B \left(\frac{\mathcal{P}_{t-1}^B}{\mathcal{P}_{t-1}^B + \mathcal{P}_{t-1}^R}\right)(1 - F_X(k_1)) + (1 - \mathcal{P}_{t-1}^B)\left(\frac{1 - \mathcal{P}_{t-1}^B}{2 - \mathcal{P}_{t-1}^B - \mathcal{P}_{t-1}^R}\right)(1 - F_X(k_1)) = q_{t-1}(1 - F_X(k_1)),$$

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.

More interestingly, a homogeneous match results in the following monotonicity relationship: For each pair $\omega_2^S, \widehat{\omega}_2^S$ satisfying $\pi_2(\widehat{\omega}_2^S) > \pi_2(\omega_2^S)$, we have $k_t(\widehat{\omega}_2^S) < k_t(\omega_2^S)$.[12]

Proposition 1 *Suppose (A1)-(A4). 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 intuitive. If a B member observes Invest from another B member when they were young, his updated belief that the B group is more likely to have a good distribution results in a lower stage game threshold when old, $k_t^S(I)$ – a higher probability of choosing Invest – which yields the first inequality, and the same argument works for the second inequality. For homogeneous group matches among old members, the history of play in interactions with members of *the other* group in youth is irrelevant, but this will no longer hold in the case of heterogeneous matches as shown in the next subsection.

4.2 Heterogeneous match

In the case where 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 (5) – the old player's beliefs about the probability that his matched partner will choose No Invest, $\Pr(N|\omega_2^A, \mathbf{k}_t^A)$, play an essential role. However, in this case, the probability 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 No Invest given each history. That is, unlike in the homogeneous match case, histories from both the same or 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, *heterogeneous* matches, can be illustrated as follows. Consider a B player who is matched with a member of the R group when old. If the B player was matched with a member of his own B group in youth and experienced a good "Invest" outcome, then the B player thinks it more likely that his old R member match may also have had a good experience with the B group in youth, which makes the B player more likely to choose Invest in the old age match with the R member. Suppose, instead, that the B player was matched with a member of the R group in youth and also experienced a good Invest decision. This also 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 the R group. As we show below in Proposition 2, the latter effect is stronger than the former so that the 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 2).

[12]Note that to simplify the notations in both this 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 exact value of $k_t(\omega_2^S)$ changes as q_{t-1} changes. Nonetheless, the monotonicity holds regardless of $q_{t-1} > 1/2$. The role of q_{t-1} becomes apparent in Proposition 3 and especially in dynamics in Section 6.

To analyze it formally, let an old player's history for a heterogeneous match be denoted by $\omega_2^A \in \Omega_2^A \equiv \{I|\emptyset, \emptyset|I, \emptyset|N, N|\emptyset\}$. 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 the B player's perspective, $\emptyset|I$ indicates that previously, the Blue (B) member was not matched with a Red (R) member, instead observing Invest from another B member. The corresponding threshold is written as $k_t^A(\omega_2)$ for $\omega_2^A \in \Omega_2^A$. A heterogeneous match equilibrium profile of the old player's thresholds is written as $k_t^A \equiv \big(k_t^A(\omega_2^A)\big)_{\{\omega_2^A \in \Omega_2^A\}}$, and a profile including the youthful k_1 threshold is written as $\mathbf{k}_t^A \equiv (k_1, k_t^A)$.

Thus, an old player in $m = A$ obtains an expected payoff from Invest equal to

$$U_t^A(\theta, \omega_2^A, \mathbf{k}_t^A) = d(\theta) - \Pr(N|\omega_2^A, \mathbf{k}_t^A). \tag{9}$$

Note that in contrast to the homogeneous case in (5), even *conditional on* the matched R group distribution $r = X^R$ or $r = Y^R$, the uncertainty with respect to the player's own B group still remains. This can be captured by the following probabilities:

$$\begin{aligned} p_2(I|\omega_2^B) &\equiv \pi_2(\omega_2^B)(1 - F_X(k_1)) + (1 - \pi_2(\omega_2^B))(1 - F_Y(k_1)), \\ p_2(N|\omega_2^B) &\equiv \pi_2(\omega_2^B)F_X(k_1) + (1 - \pi_2(\omega_2^B))F_Y(k_1). \end{aligned} \tag{10}$$

That is, $p_2(I|\omega_2^B)$ (resp. $p_2(N|\omega_2^B)$) is an old B member's belief that his matched R player, in youth, observed Invest (resp. No Invest) from another B player previously.

A heterogeneous match equilibrium is defined such that for each $\ell \in \{E, W\}$ and every $\omega_2^A \in \Omega_2^A$,

$$d(k_t^A(\omega_2^A)) = \pi_2(\omega_2^R)X_t^A(\mathbf{k}_t^A, \omega_2^B) + (1 - \pi_2(\omega_2^R))Y_t^A(\mathbf{k}_t^A, \omega_2^B). \tag{11}$$

One has to interpret equation (11) carefully. On the left-hand side (LHS), $k_t^A(\omega_2^A)$ is from an old B player's point of view, that is, $\omega_2^A = \omega_2^R|\omega_2^B$, which is incorporated into $\pi_2(\omega_2^R)$ and $\pi_2(\omega_2^B)$ through $p_2(\cdot|\omega_2^B)$ in $X_t^A(\mathbf{k}_t^A, \omega_2^B)$ and $Y_t^A(\mathbf{k}_t^A, \omega_2^B)$ 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 $\omega_2^B|\omega_2^R$, *i.e.*, from R's perspective. Extending a homogeneous match equilibrium in (8), 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

$$\mathbf{\Phi}_t^A(k_t^A) \equiv \Big(\Phi_t^A(k_t^A, \omega_2^A)\Big)_{\{\omega_2^A \in \Omega_2^A\}}, \tag{12}$$

where $\Phi_t^A(k_t^A, \omega_2^A) \equiv d^{-1}\big(\pi_2(\omega_2^R)X_t^A(\mathbf{k}_t^A, \omega_2^B) + (1 - \pi_2(\omega_2^R))Y_t^A(\mathbf{k}_t^A, \omega_2^B)\big)$ (see the proof of Lemma 2 for more details).

Unlike $X_t^S(\mathbf{k}_t^S)$ and $Y_t^S(\mathbf{k}_t^S)$ in the homogeneous match case, since $X_t^A(\mathbf{k}_t^A, \omega_2^B)$ and $Y_t^A(\mathbf{k}_t^A, \omega_2^B)$ depend on ω_2^B, it is not straightforward to find a monotonicity result for a profile of the old player's thresholds k_t^A. 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 1, monotonicity between $k_t^A(I|\emptyset)$ and $k_t^A(\emptyset|I)$ or that between $k_t^A(N|\emptyset)$ and $k_t^A(\emptyset|N)$ demands a whole new approach. The lemma below is the first step in this direction.

Lemma 2 *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) $d(k_t^A(I|\emptyset)) - d(k_t^A(\emptyset|I)) < q_{t-1}(\pi_2(I^R) - \pi_2(\emptyset^R))(F_Y(k_1) - F_X(k_1))\left[F_r(k_t^A(\emptyset|I)) - F_r(k_t^A(\emptyset|N))\right]$,

(ii) $d(k_t^A(N|\emptyset))-d(k_t^A(\emptyset|N)) > q_{t-1}(\pi_2(N^R)-\pi_2(\emptyset^R))(F_Y(k_1)-F_X(k_1))\left[F_r(k_t^A(\emptyset|I)) - F_r(k_t^A(\emptyset|N))\right]$,

(iii) $d(k_t^A(\emptyset|I))-d(k_t^A(\emptyset|N)) = \frac{(1-q_{t-1})}{2}(\pi_2(I^B)-\pi_2(N^B))[F_Y(k_1)-F_X(k_1)]\left[\begin{array}{l} F_X(k_t^A(I|\emptyset)) - F_X(k_t^A(N|\emptyset)) \\ +F_Y(k_t^A(I|\emptyset)) - F_Y(k_t^A(N|\emptyset)) \end{array}\right]$.

To establish a monotonicity result for a heterogeneous match, we first prove that $k_t^A(N|\emptyset) > k_t^A(I|\emptyset)$ in the following proposition. Once this is shown, then, Lemma 2 (iii) implies that $k_t^A(\emptyset|N) > k_t^A(\emptyset|I)$, which in turn can be incorporated into Lemma 2 (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.

Proposition 2 *Suppose (A1)-(A4). 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, but 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 No Invest from another B player when young than if he observed No Invest from an R player when young. In sum, the experience with a young R player in the past *reinforces* his investment decision with a matched old R player in either direction, compared with the same experience with a young B player.

Interestingly, if the B player has no youthful history with an R player then $k_t^A(\emptyset|I) < k_t^A(\emptyset|N)$. The intuition is that if the old B player met a B player in youth who chose to Invest (No Invest), then the old B player believes that his current matched R partner is more likely to have also had a good (bad) experience with a B player when he was young, making the old B player more (less) likely to choose Invest in the match with the old R player.

5 Location stage equilibrium

In this section, we use the investment stage equilibrium characterizations from homogeneous and heterogeneous matches to determine a location stage equilibrium. That is, given *beliefs* about which location has more of the same group members, each player optimally chooses whether they want to locate in the East or the West, comparing the two future payoffs when old: one from meeting with the same group member and another from meeting with a different group member. For this analysis, we introduce the notation $\omega_2 = (\omega_2^B, \omega_2^R)$ with a fixed order, unlike ω_2^A. Recall that in a heterogeneous match of the investment stage, an equilibrium is properly defined only when the history ω_2^A takes the group player's own point of view with the observation from the matched partner's group – a different group – first. By contrast, a location stage equilibrium considers both homogeneous and heterogeneous matches, which requires a more "neutral" notation.

Let $P_t^B(\omega_2, \ell_t^B)$ denote a player's beliefs given history ω_2 about the proportion of B players located in E among B members in period t, and $P_t^R(\omega_2, \ell_t^R)$ be his belief given history ω_2 about the proportion of R players located in E among R members in period t. In this section, we consider the

case where $P_t^B(\omega_2, \ell_t^B) \neq P_t^R(\omega_2, \ell_t^R)$, and we deal with the remaining case in the next section.[13] The location decision made by a player who is a member of group $g \in \{B, R\}$ and who resides in either location is given by a mapping

$$\ell_t^g : \Omega_2 \to \{E, W\}, \tag{13}$$

where $\Omega_2 \equiv \{(I, \emptyset), (\emptyset, N), (N, \emptyset), (\emptyset, I)\}$. We consider a B player without loss of generality. If a B player chooses E, he obtains the expected payoff

$$V_t^B(E, \theta, \omega_2, \ell_t) = \frac{P_t^B(\omega_2, \ell_t^B)}{P_t^B(\omega_2, \ell_t^B) + P_t^R(\omega_2, \ell_t^R)} U_t^S(\theta, \omega_2^S, \mathbf{k}_t^S) + \frac{P_t^R(\omega_2, \ell_t^R)}{P_t^B(\omega_2, \ell_t^B) + P_t^R(\omega_2, \ell_t^R)} U_t^A(\theta, \omega_2^A, \mathbf{k}_t^A),$$

where $\ell_t \equiv (\ell_t^B, \ell_t^R)$. If, on the other hand, the B player chooses W, he obtains the expected payoff

$$V_t^B(W, \theta, \omega_2, \ell_t) = \frac{1 - P_t^B(\omega_2, \ell_t^B)}{2 - P_t^B(\omega_2, \ell_t^B) - P_t^R(\omega_2, \ell_t^R)} U_t^S(\theta, \omega_2^S, \mathbf{k}_t^S) + \frac{1 - P_t^R(\omega_2, \ell_t^R)}{2 - P_t^B(\omega_2, \ell_t^B) - P_t^R(\omega_2, \ell_t^R)} U_t^A(\theta, \omega_2^A, \mathbf{k}_t^A).$$

We now provide a formal definition for a location equilibrium $\ell_t = (\ell_t^B, \ell_t^R)$ at each period t.

Definition 1 *ℓ_t is said to be a location (Bayesian) equilibrium at $t = 1, 2, ...$ if for each $g \in \{B, R\}$ and every $(\theta, \omega_2) \in \Theta \times \Omega_2$,*

$$\begin{aligned} &(i)\ V_t^g(\ell_t^g(\omega_2), \theta, \omega_2, \ell_t) \geq V_t^g(\ell, \theta, \omega_2, \ell_t) \text{ for all } \ell \in \{E, W\}. \\ &(ii)\ P_t^B(\omega_2, \ell_t^B) = \mathbb{E}\big[\mathbf{1}_{\{\ell_t^B(\widetilde{\omega}_2) = E\}} \mid \omega_2\big] \text{ and } P_t^R(\omega_2, \ell_t^R) = \mathbb{E}\big[\mathbf{1}_{\{\ell_t^R(\widetilde{\omega}_2) = E\}} \mid \omega_2\big]. \end{aligned}$$

The first condition (i) addresses optimality and the second condition (ii) addresses *consistency* such that each player's expectations about the other players' location strategies are correct. For example, consider a player with $\omega_2 = (I, \emptyset)$ who forms beliefs about $P_t^B(\omega_2, \ell_t^B)$ by expecting $\ell_t^B(\widetilde{\omega}_2)$ for all $\widetilde{\omega}_2 \in \Omega_2$, where $\widetilde{\omega}_2$ denotes a history that the other player *can* have, and in equilibrium, the expectations must be correct in the sense that they are identical to the actual equilibrium choices of players with other histories. However, this does *not* mean that $P_t^B(\omega_2, \ell_t^B) = \mathcal{P}_t^B$ from (4); 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.

Specifically, the beliefs, $P_t^B(\omega_2, \ell_t^B)$ and $P_t^R(\omega_2, \ell_t^R)$, can be derived such that

$$\begin{aligned} P_t^B(\omega_2, \ell_t^B) = {} & q_{t-1} p(I|\omega_2^B) \mathbf{1}_{\{\ell_t^B(I,\emptyset) = E\}} + (1 - q_{t-1}) p(N|\omega_2^R) \mathbf{1}_{\{\ell_t^B(\emptyset,N) = E\}} \\ & + q_{t-1} p(N|\omega_2^B) \mathbf{1}_{\{\ell_t^B(N,\emptyset) = E\}} + (1 - q_{t-1}) p(I|\omega_2^R) \mathbf{1}_{\{\ell_t^B(\emptyset,I) = E\}}, \end{aligned} \tag{14}$$

and

$$\begin{aligned} P_t^R(\omega_2, \ell_t^R) = {} & (1 - q_{t-1}) p(I|\omega_2^B) \mathbf{1}_{\{\ell_t^R(I,\emptyset) = E\}} + q_{t-1} p(N|\omega_2^R) \mathbf{1}_{\{\ell_t^R(\emptyset,N) = E\}} \\ & + (1 - q_{t-1}) p(N|\omega_2^B) \mathbf{1}_{\{\ell_t^R(N,\emptyset) = E\}} + q_{t-1} p(I|\omega_2^R) \mathbf{1}_{\{\ell_t^R(\emptyset,I) = E\}}. \end{aligned} \tag{15}$$

[13] If $P_t^B(\omega_2, \ell_t^B) = P_t^R(\omega_2, \ell_t^R)$, then players are indifferent between moving to E or W, which means that for the same history, some portion of members of the same group with that history can choose one place, whereas the remaining portion can choose the other place to yield the equality. However, if not, a set of location equilibria reduces to a simple class, as will be shown subsequently.

Consider (14) and further, for instance, $\widetilde{\omega}_2 = (I, \emptyset)$ as a history that the other player can have among the four possible histories. Then, $q_{t-1}p(I|\omega_2^B)\mathbf{1}_{\{\ell_t^B(I,\emptyset)=E\}}$ means that a player with history ω_2 reasons, *based on his own experience* ω_2, that with probability q_{t-1}, an arbitrary B player met with a member of his own group and with probability $p(I|\omega_2^B)$, this other player observed I, and if a B player with $\widetilde{\omega}_2 = (I, \emptyset)$ moves to E, *i.e.*, $\mathbf{1}_{\{\ell_t^B(I,\emptyset)=E\}} = 1$, then that population proportion must be *counted* for $P_t^B(\omega_2, \ell_t^B)$.[14] The reasoning relies on $p(I|\omega_2^g)$ and $p(N|\omega_2^g)$ in (10) from $\omega_2 = (\omega_2^B, \omega_2^R)$. The following lemma shows that the belief difference between $P_t^B(\omega_2, \ell_t^B)$ and $P_t^R(\omega_2, \ell_t^R)$ plays a critical role in both group members' location decisions, which is denoted by

$$\Delta P_t(\omega_2, \ell_t) \equiv P_t^B(\omega_2, \ell_t^B) - P_t^R(\omega_2, \ell_t^R). \tag{16}$$

Lemma 3 provides an important intermediate step for how one can actually find an equilibrium using the location equilibrium definition in Definition 1. In addition, it shows that there is no role for a player's intrinsic type θ in the location stage decision; consistent with the statistical discrimination perspective, the only private information that matters for the location equilibrium is the *history* of observations ω_2.

Lemma 3 *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*

$$U_t^S(\theta, \omega_2^S, \mathbf{k}_t^S) - U_t^A(\theta, \omega_2^A, \mathbf{k}_t^A) = d(k_t^A(\omega_2^A)) - d(k_t^S(\omega_2^S)),$$

and in equilibrium, the optimal location decisions are given as follows.

(i) any B member with ω_2 chooses E if $\Delta P_t(\omega_2, \ell_t)\left[d(k_t^A(\omega_2^A)) - d(k_t^S(\omega_2^S))\right] > 0$.

(ii) any R member with ω_2 chooses E if $-\Delta P_t(\omega_2, \ell_t)\left[d(k_t^A(\omega_2^A)) - d(k_t^S(\omega_2^S))\right] > 0$.

The intuition behind this result appears rather straightforward at first in the sense that players will move toward the location in which they expect to earn a higher payoff. Yet, it also reveals the delicate nature of the problem. To see that, let us delve further into the result, observing that the condition above can be further divided into two components: $d(k_t^A(\omega_2^A)) - d(k_t^S(\omega_2^S))$ and $\Delta P_t(\omega_2, \ell_t)$. Now consider a particular history $\omega_2 = (\omega_2^B, \omega_2^R)$ from among four histories in Ω. Then, first, we need to determine whether ω_2, by incorporating ω_2 into ω_2^S and ω_2^A, yields $d(k_t^A(\omega_2^A)) > d(k_t^S(\omega_2^S))$ or not. In other words, the first critical element is to *classify* what histories make each player expect such a "favorable" stance toward the player's 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 $\Delta P_t(\omega_2, \ell_t)$ is based on (14) and (15) that in fact contains location strategies for *all* four histories. This means that there exists a location equilibrium only when for each history $\omega_2 \in \Omega$, a player with ω_2 chooses a location that is "compatible" with his incentive to do so, provided that all other players with other histories make location choices that he exactly

[14] For $P_t^R(\omega_2, \ell_t^R)$, $(1 - q_{t-1})p(I|\omega_2^B)\mathbf{1}_{\{\ell_t^R(I,\emptyset)=E\}}$ means that a player with a history ω_2 reasons, based on his own experience ω_2, that with probability $1 - q_{t-1}$, an arbitrary R player met with a different member, a B member, and with probability $p(I|\omega_2^B)$, he observed I, and if an R player with $\widetilde{\omega}_2 = (I, \emptyset)$ moves to E, *i.e.*, $\mathbf{1}_{\{\ell_t^R(I,\emptyset)=E\}} = 1$, the portion must be counted for $P_t^R(\omega_2, \ell_t^R)$.

expects – more (less) of his own group together with a history favorable (unfavorable) toward his group. Since all the players choose their location strategies simultaneously, in a location stage equilibrium, all beliefs across the two group members with four possible histories must *clear* in the precise sense that they satisfy the consistency condition (ii) in Definition 1.

To tackle the first component, we start by defining a B player's set of histories that yield favorable and unfavorable stances toward his own group as Ω_2^{B+} and Ω_2^{B-}, respectively – recall that a lower threshold means a higher probability of choosing Invest – such that

$$\Omega_2^{B+} \equiv \left\{\omega_2 \in \Omega_2 : k_t^A(\omega_2^A) > k_t^S(\omega_2^S)\right\} \text{ and } \Omega_2^{B-} \equiv \left\{\omega_2 \in \Omega_2 : k_t^A(\omega_2^A) < k_t^S(\omega_2^S)\right\}. \tag{17}$$

The corresponding sets can be defined for R players as well. Now, to *identify* these sets reduces to comparing a profile of the old player's thresholds from a homogeneous match equilibrium in Proposition 1 with that from a heterogeneous match equilibrium in Proposition 2 in Section 4. The two profiles of thresholds are fixed points from two mappings, one in (8) and the other in (12), which in turn implies that the comparison requires a comparison between the two fixed points and thus between the two mappings. Despite challenges with respect to comparing thresholds from two different mappings – the three-dimensional homogeneous mapping $\mathbf{\Phi}_t^S : [\underline{\theta}, \overline{\theta}]^3 \to [\underline{\theta}, \overline{\theta}]^3$ in (8) and the four-dimensional heterogeneous mapping $\mathbf{\Phi}_t^A : [\underline{\theta}, \overline{\theta}]^4 \to [\underline{\theta}, \overline{\theta}]^4$ in (12) – we employ a simple but yet clever method to show it: 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 $p_2(\cdot|\omega_2^B)$ in (10) with $\lambda \in [0, 1]$ such that for the part with X_t^A,

$$\begin{aligned} p_2^X(I|\omega_2^B, \lambda) &\equiv [1 - (1 - \pi_2(\omega_2^B))\lambda](1 - F_X(k_1)) + (1 - \pi_2(\omega_2^B))\lambda(1 - F_Y(k_1)), \\ p_2^X(N|\omega_2^B, \lambda) &\equiv [1 - (1 - \pi_2(\omega_2^B))\lambda]F_X(k_1) + (1 - \pi_2(\omega_2^B))\lambda F_Y(k_1); \end{aligned} \tag{18}$$

and for the part with Y_t^A,

$$\begin{aligned} p_2^Y(I|\omega_2^B, \lambda) &\equiv \pi_2(\omega_2^B)\lambda(1 - F_X(k_1)) + [1 - \pi_2(\omega_2^B)\lambda](1 - F_Y(k_1)), \\ p_2^Y(N|\omega_2^B, \lambda) &\equiv \pi_2(\omega_2^B)\lambda F_X(k_1) + [1 - \pi_2(\omega_2^B)\lambda]F_Y(k_1). \end{aligned} \tag{19}$$

If we replace (10) by (18) and (19) – we now have different values of p_2 depending on X and Y to construct the auxiliary mapping – then, as one can find in the proof of Proposition 3, it connects the two mappings: If $\lambda = 0$, we have the symmetric model, whereas if $\lambda = 1$, we have the asymmetric model. 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. Note that if $\lambda = 0$, $\widehat{\Phi}_t(\widehat{k}_t^\lambda, \emptyset|I, \lambda) = \widehat{\Phi}_t(\widehat{k}_t^\lambda, \emptyset|N, \lambda)$, so *effectively*, there are three functions for $\lambda = 0$, which is the homogeneous mapping. The auxiliary mapping's equilibrium profile of the old player's thresholds is written as $\widehat{k}_t^\lambda \equiv \big(\widehat{k}_t^\lambda(\omega_2^A)\big)\big)_{\{\omega_2^A \in \Omega_2^A\}}$, and a profile including the youthful k_1 threshold is written as $\widehat{\mathbf{k}}_t^\lambda \equiv (k_1, \widehat{k}_t^\lambda)$. In order to make the auxiliary mapping increase in λ, we need an additional condition. Since λ is not germane to the model, we make an additional assumption: given $m \equiv \max\left\{\frac{F_Y(k_1)}{F_X(k_1)}, \frac{1-F_X(k_1)}{1-F_Y(k_1)}\right\}$,

(A5) For each $\lambda > 0$, $F_X(\widehat{k}_t^\lambda(N|\emptyset))) - F_X(\widehat{k}_t^\lambda(I|\emptyset)) \geq m\big[F_Y(\widehat{k}_t^\lambda(N|\emptyset)) - F_Y(\widehat{k}_t^\lambda(I|\emptyset))\big]$.

This assumption requires that for increases in the probability of choosing No Invest, from $\widehat{k}_t^\lambda(I|\emptyset))$ to $\widehat{k}_t^\lambda(N|\emptyset))$, F_X dominates F_Y for a weight $m > 1$.[15] Assumption (A5) is satisfied for a large

[15]The definition m is also related to the uniqueness in Section 7.

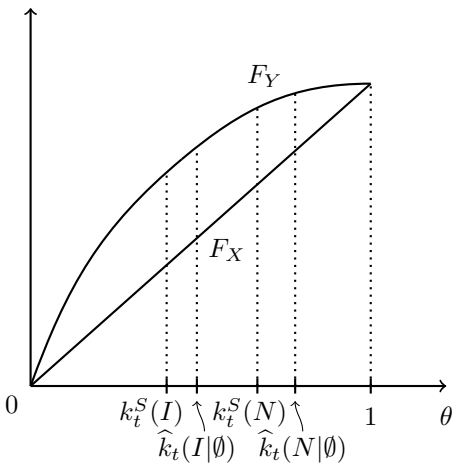

Figure 3: F_k satisfying (A5)

class of distributions, and here we provide two such cases. First, initially, suppose $F_X(k_t^S(N)) - F_X(k_t^S(I)) \geq m\big[F_Y(k_t^S(N)) - F_Y(k_t^S(I))\big]$. If F_X is convex but F_Y is concave on the effective support $[k_X, k_Y]$, then (A5) holds.[16] This case is illustrated in Figure 3 for an example with $F_X(\theta) = \theta$ and $F_Y(\theta) = \frac{1-e^{-\lambda\theta}}{1-e^{-\lambda}}$ for $\theta \in [0,1]$. Second, one can find that a sufficient condition for (A5) is $\frac{f_X(\theta)}{f_Y(\theta)} \geq m$, where $\frac{f_X(\theta)}{f_Y(\theta)}$ is the likelihood ratio, which can be called the *bounded* likelihood ratio condition. 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 *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. 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.[17]

Proposition 3 *(Comparison between two types of matches) Suppose (A1)-(A5) and* $\Delta P_t(\omega_2, \ell_t) \neq 0$. *Then, for each* $t = 1, 2...$, *in equilibrium, for an initial population difference greater than a critical value, the relationship between the thresholds in a homogeneous match and the thresholds in a heterogeneous match is given as follows.*

(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) from the monotone comparative statics, but the second set of results in (ii) requires that the heterogeneous equilibrium thresholds *do not* increase too much (see Appendix for the formal proof). In particular, it can be readily shown that as the society

[16] For each λ, $\frac{F_X(\widehat{k}_t^\lambda(N|\emptyset))) - F_X(\widehat{k}_t^\lambda(I|\emptyset))}{\widehat{k}_t^\lambda(N|\emptyset)) - \widehat{k}_t^\lambda(I|\emptyset))} \geq \frac{F_X(k_t^S(N)) - F_X(k_t^S(I))}{k_t^S(N) - k_t^S(I)}$ and $\frac{F_Y(k_t^S(N)) - F_Y(k_t^S(I))}{k_t^S(N) - k_t^S(I)} \geq \frac{F_Y(\widehat{k}_t^\lambda(N|\emptyset))) - F_Y(\widehat{k}_t^\lambda(I|\emptyset)))}{\widehat{k}_t^\lambda(N|\emptyset)) - \widehat{k}_t^\lambda(I|\emptyset))}$.

[17] A condition slightly stronger than (A3) can guarantee uniqueness for k_t^S, which is discussed in Section 7.

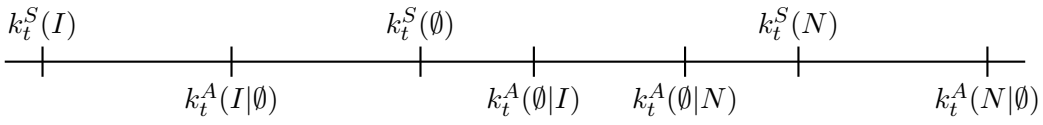

Figure 4: Comparison between two matches

approaches the perfectly polarized state, that is, for q_{t-1} close 1, the two equilibrium thresholds, the fixed point in (8) and the one in (12), are sufficiently close to each other that they satisfy the second set of results.

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 3. In the proof of Proposition 7, we will use the maximum of these two critical values which will insure that Proposition continues to hold.

Summarizing the discussion to this point, Figure 4 shows the relationships between homogeneous and heterogeneous matches arising out of Propositions 1-3.

By Proposition 3, the set of histories favorable and unfavorable to playing the game with own group members are given by:

$$\begin{aligned} \Omega_2^{B+} &= \{(I,\emptyset),(\emptyset,N)\} \text{ and } \Omega_2^{B-} = \{(N,\emptyset),(\emptyset,I)\}, \\ \Omega_2^{R+} &= \{(\emptyset,I),(N,\emptyset)\} \text{ and } \Omega_2^{R-} = \{(\emptyset,N),(I,\emptyset)\}. \end{aligned} \tag{20}$$

The four comparisons reveal how the same histories can lead to a differences 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. No Invest), *i.e.* $\omega_2 = (I,\emptyset)$ (resp. $\omega_2 = (N,\emptyset)$), by Lemma 3, a matching with a member of the same B group when old yields a higher (resp. lower) future expected payoff than a matching with an R player. 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 intuition applies to the cases of matching with a member of the other group in the past, *i.e.*, $\omega_2 = (\emptyset, I)$ or $\omega_2 = (\emptyset, N)$.[18]

Now, for the second critical element $\Delta P_t(\omega_2, \ell_t)$, observing that for each $\omega_2 \in \Omega$, we have $\ell_t^B(\omega_2) = \ell_t^R(\omega_2)$ as the first result of the following lemma. Then, from (14) and (15),

$$\begin{aligned} \Delta P_t(\omega_2,\ell_t) =& (2q_{t-1}-1)[\mathbf{1}_{\{\ell_t^g(I,\emptyset)=E\}} - \mathbf{1}_{\{\ell_t^g(\emptyset,I)=E\}}] \\ &+ (2q_{t-1}-1)p(N|\omega_2^B)[\mathbf{1}_{\{\ell_t^g(N,\emptyset)=E\}} - \mathbf{1}_{\{\ell_t^g(I,\emptyset)=E\}}] \\ &+ (2q_{t-1}-1)p(N|\omega_2^R)[\mathbf{1}_{\{\ell_t^g(\emptyset,I)=E\}} - \mathbf{1}_{\{\ell_t^g(\emptyset,N)=E\}}]. \end{aligned} \tag{21}$$

Each player has to anticipate whether there will be more B members in the East (or not), *i.e.*, $\Delta P_t(\omega_2, \ell_t) > 0\,(< 0)$, to make his location decision. The expectation is based on whether a history $\widetilde{\omega} \in \Omega$ the other player can have is favorable toward that player's own group or not – that is,

[18]Despite the equality $\Omega_2^{B+} = \Omega_2^{R-}$, one needs to be careful about interpretations since $(I,\emptyset)$ in Ω_2^{B+} means that a B player observed Invest from a member of the same group, whereas $(I,\emptyset)$ in Ω_2^{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_2^{R-} is identical to Ω_2^{B-}.

	$\Delta P_t(\omega_2, \ell_t) > 0$ for all ω_2	$\Delta P_t(\omega_2, \ell_t) < 0$ for all ω_2
$\omega_2 \in \Omega_2^{B+} = \Omega_2^{R-}$	B (R) chooses E	B (R) chooses W
$\omega_2 \in \Omega_2^{B-} = \Omega_2^{R+}$	B (R) chooses W	B (R) chooses E

Table 2: The location stage equilibrium

each difference between two indicator functions in (21) has a form of $\mathbf{1}_{\{\omega_2 \in \Omega_2^{B+}\}} - \mathbf{1}_{\{\omega_2 \in \Omega_2^{B-}\}}$ (or negative of this), in light of the sets we identified in (20). In particular, as discussed earlier, in doing so, the probability of observing a history is based on his personal experience with $\omega_2 = (\omega_2^B, \omega_2^R)$ incorporated into $p(N|\omega_2^B)$ and $p(N|\omega_2^R)$ above.

There are 2^8 strategies to consider given that each group member has four different histories $\Omega_2^{g+}, \Omega_2^{g-}$ in (20) with two locations; in other words, each player with a history favorable toward his own group can choose E if there are more B players in E, that is, if $\Delta P_t(\omega_2, \ell_t) > 0$ but can choose W if there are more B players in W, that is, if $\Delta P_t(\omega_2, \ell_t) < 0$. Despite the large number of strategies, the following lemma shows that we can reduce the number of possible strategies for a location equilibrium significantly.

Lemma 4 *Suppose (A1)-(A5). Then for each $g \in \{B, R\}$ and $\Delta P_t(\omega_2, \ell_t) \neq 0$, a location equilibrium ℓ_t given $\Omega_2^{g+}, \Omega_2^{g-}$ satisfies the following properties.*

(i) For each $\omega_2 \in \Omega$, we have $\ell_t^B(\omega_2) = \ell_t^R(\omega_2)$.

(ii) For each $\omega_2 \neq \omega_2' \in \Omega_2^{g+}$, $\ell_t^g(\omega_2) = \ell_t^g(\omega_2')$ and $\omega_2 \neq \omega_2' \in \Omega_2^{g-}$, $\ell_t^g(\omega_2) = \ell_t^g(\omega_2')$.

(iii) For each $\omega_2 \in \Omega_2^{g+}$, $\omega_2' \in \Omega_2^{g-}$, $\ell_t^g(\omega) \neq \ell_t^g(\omega')$.

By Lemma 4, if there exists a location equilibrium, then by its very nature, it will be a *binary splitting* equilibrium such that for each $g \in \{B, R\}$,

$$\forall \omega_2 \in \Omega_2^{B+} = \Omega_2^{R-},\ \omega_2' \in \Omega_2^{B-} = \Omega_2^{R+},\ \begin{cases} \ell_t^g(\omega_2) = E, \ell_t^g(\omega_2') = W & \text{if } \forall \omega_2,\ \Delta P_t(\omega_2, \ell_t) > 0, \\ \ell_t^g(\omega_2) = W, \ell_t^g(\omega_2') = E & \text{if } \forall \omega_2,\ \Delta P_t(\omega_2, \ell_t) < 0, \end{cases} \tag{22}$$

which is laid out in Table 2, where players with any two different histories $\omega_2 \neq \omega_2'$ share the *same beliefs about the sign* of $\Delta P_t(\omega_2, \ell_t)$, either $\Delta P_t(\omega_2, \ell_t) > 0$ or $\Delta P_t(\omega_2, \ell_t) < 0$, as a consequence of Lemma 4 (ii) & (iii). The location equilibrium is binary splitting in the following sense. If $P_t^B(\omega_2, \ell_t) > P_t^R(\omega_2, \ell_t)$, and $\omega_2 \in \Omega_2^{B+}$, then a B player chooses E over W, expecting a higher likelihood to meet another B player based on a *good* experience with the same group member, whereas if $P_t^B(\omega_2, \ell_t) > P_t^R(\omega_2, \ell_t)$ and $\omega_2 \in \Omega_2^{B-}$, then a B player chooses W over E, expecting a higher likelihood to meet another R player based on a *bad* experience with the same group member. The analysis can be appropriately modified for an R player.

Then, by incorporating the binary splitting location equilibrium into (21), we have (the detailed procedure is relegated to the proof of Proposition 4):

$$\Delta P_t(\omega_2, \ell_t) = \begin{cases} (2q_{t-1} - 1)[1 - p_2(N|\omega_2^B) - p_2(N|\omega_2^R)] & \text{if } \Delta P_t(\omega_2, \ell_t) > 0, \\ -(2q_{t-1} - 1)[1 - p_2(N|\omega_2^B) - p_2(N|\omega_2^R)] & \text{if } \Delta P_t(\omega_2, \ell_t) < 0. \end{cases} \tag{23}$$

By the natural law of likes meeting likes (Lemma 1), the same group matching probability is higher; $2q_{t-1} - 1 > 0$. Hence, if players believe $\Delta P_t(\omega_2, \ell_t) > 0$ (resp. < 0), then it *actually* arises from

$\ell_t^g(\omega) = E, \ell_t^g(\omega') = W$ (resp. $\ell_t^g(\omega) = W, \ell_t^g(\omega') = E$) for $\omega_2 \in \Omega_2^{B+} = \Omega_2^{R-}$, $\omega_2' \in \Omega_2^{B-} = \Omega_2^{R+}$ when the second term satisfies $1 - p_2(N|\omega_2^B) - p_2(N|\omega_2^R) > 0$.

To provide a tight connection between the investment stage equilibrium when old in Section 4 and the dynamics in Section 6 as well as the location stage equilibrium in this section, we examine the formula (23) further, in particular by rewriting its first derivation (21) as

$$\begin{aligned}\Delta P_t(\omega_2, \ell_t) =&(2q_{t-1} - 1)\big[p(I|\omega_2^B)\mathbf{1}_{\{\ell_t^g(I,\phi)=E\}} - p(N|\omega_2^R)\mathbf{1}_{\{\ell_t^g(\phi,N)=E\}}\big] \\ &- (2q_{t-1} - 1)\big[p(I|\omega_2^R)\mathbf{1}_{\{\ell_t^g(\phi,I)=E\}} - p(N|\omega_2^B)\mathbf{1}_{\{\ell_t^g(N,\phi)=E\}}\big].\end{aligned}$$

Suppose that $\Delta P_t(\omega_2, \ell_t) > 0$; that is, there are more B members in the East in equilibrium. Consider $(I, \emptyset)$, which belongs to Ω_2^{B+} (but not to Ω_2^{B-}). That is, the case of a favorable history towards his own group for a B member but an unfavorable history toward his own group for a R member. As such, 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 his own group members when old, whereas a R member with that history finds it optimal to go to E to meet with members of the other group. As a result, the former has the effect of widening the polarization but the latter has the effect of reducing it. The overall effect is positive due to the law of likes meeting likes: The former match (B, B) arises with a higher probability in youth. This is the part $(2q_{t-1} - 1)p(I|\omega_2^B)\mathbf{1}_{\{\ell_t^g(I,\phi)=E\}}$. Now, consider another same kind of history $(\emptyset, N)$ – which also belongs to Ω_2^{B+}. In this case, however, the overall effect is now negative due to the same law of likes meeting likes: The latter match (R, R) arises with a higher probability in youth. If the belief $\Delta P_t(\omega_2, \ell_t)$ arises in equilibrium, then

$$\Delta P_t(\omega_2, \ell_t) - \Delta P_{t-1}(\omega_2, \ell_{t-1}) = (2q_{t-1} - 1)[p(I|\omega_2^B) - p(N|\omega_2^R)] - \Delta P_{t-1}(\omega_2, \ell_{t-1}).$$

The real dynamical system in the next section is given from the point of view of the modeller who *knows* the real F_B and F_R. Therefore, we can translate the above difference into $\Delta\mathcal{P}_t - \Delta\mathcal{P}_{t-1}$ from $\Delta\mathcal{P}_{t-1} \equiv \mathcal{P}_{t-1}^B - \mathcal{P}_{t-1}^R$ which is the difference in the actual moving probabilities with which each group moves to the East, where $\mathcal{P}_{t-1}^B$ and $\mathcal{P}_{t-1}^R$ are from Section 3. Hence, if q_{t-1} is small, the first term becomes small, so the difference is negative and the system converges to a completely mixed state. On the other hand, if q_{t-1} is sufficiently high, the difference is positive and the dynamics converge to 1, the case of complete polarization. Notice that the investment stage equilibrium of Section 4, through the location equilibrium, provides the foundation for the convergence results obtained in Section 6.

In the beliefs, as well as in the subsequent "real" dynamics, the *sum of the beliefs about the probabilities* of observing No Invest plays a key role in (23), which is defined as

$$N_2(\omega_2) \equiv p_2(N|\omega_2^B) + p_2(N|\omega_2^R), \tag{24}$$

where the vector's order does not matter, *e.g.*, $N_2(\omega_2) = N_2(\omega_2')$ for $\omega_2 = (I, \emptyset)$, $\omega' = (\emptyset, I)$. By the formula in (23) combined with Lemma 1, we can identify the necessary and sufficient condition for the existence of a location equilibrium.

Proposition 4 *Suppose (A1)-(A5). Then, any location equilibrium is a binary splitting location equilibrium if and only if* $N_2(N, \emptyset) < 1$.

Henceforth, we restrict location equilibrium to be binary splitting equilibria and for brevity we will just refer to these as "location equilibria." In the next section we incorporate location equilibrium into a dynamical system.

6 Matching dynamics and polarization

In this section, we study the matching dynamics of the system over time. We do this from the perspective of an outside theorist who perfectly knows the distributions F_B and F_R. For the remainder of this section we assume that some proportion of each group makes location decisions exogenously rather than endogenously.

The reason for that assumption is twofold. First, given (23), for any q_{t-1} satisfying Lemma 1, both $\Delta P_t(\omega_2, \ell_t) > 0$ and $\Delta P_t(\omega_2, \ell_t) < 0$ 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 period t can arise in both ways: There will be more B or more R in the East in period t. Further, the equilibrium can change from one period to next, with no reason provided; $\Delta P_t(\omega_2, \ell_t) > 0$ but $\Delta P_{t+1}(\omega_2, \ell_{t+1}) < 0$. However, by allowing for some proportion of exogenous location choices, we can construct a Markov location strategy such that each player's period t location strategy depends only on the period $t-1$ actual population difference which is a more reasonable and robust restriction. The second reason for introducing non-strategic part is that it enables us to show the possibility of both completely mixed and polarized outcomes as we explain in detail below.

Specifically, suppose a proportion $\alpha \in (0,1)$ of each group $g \in \{B, R\}$ make rational endogenous location choices as in the previous section, but the remaining $1-\alpha$ proportion is under some external influence (*e.g.*, from social media or social networks). One way to interpret this exogenous force is that it is the degree of a *systematic polarization* in the society. That is, if there are *more* B members in the East, *i.e.*, if $\Delta\mathcal{P}_{t-1} > 0$, then we suppose that an additional $\epsilon\Delta\mathcal{P}_{t-1}$, for some given $\epsilon > 0$, *do not change* their location (they stay 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 $\epsilon\Delta\mathcal{P}_{t-1}$ *change* their location to the West. Hence, for $1-\alpha$ proportion, $\mathcal{P}^B_{t-1} + \epsilon\Delta\mathcal{P}_{t-1}$ can be interpreted as the B group's *inertia coupled with an amplifying effect* from the two possible group size differences in the East. An identical external force is applied to R group members.

Now, with the addition of exogenous moves and by incorporating the binary splitting location equilibria in (22) into (14) and (15), we have[19]

$$P^B_t(\omega_2, \ell^B_t) = \begin{cases} \alpha[q_{t-1}p(I|\omega^B_2) + (1-q_{t-1})p(N|\omega^R_2)] + (1-\alpha)(\mathcal{P}^B_{t-1} + \epsilon\Delta\mathcal{P}_{t-1}) & \text{if } \Delta P_t(\omega_2, \ell_t) > 0, \\ \alpha[q_{t-1}p(N|\omega^B_2) + (1-q_{t-1})p(I|\omega^R_2)] + (1-\alpha)(\mathcal{P}^B_{t-1} + \epsilon\Delta\mathcal{P}_{t-1}) & \text{if } \Delta P_t(\omega_2, \ell_t) < 0, \end{cases}$$

and

$$P^R_t(\omega_2, \ell^R_t) = \begin{cases} \alpha[(1-q_{t-1})p(I|\omega^B_2) + q_{t-1}p(N|\omega^R_2)] + (1-\alpha)(\mathcal{P}^R_{t-1} - \epsilon\Delta\mathcal{P}_{t-1}) & \text{if } \Delta P_t(\omega_2, \ell_t) > 0, \\ \alpha[(1-q_{t-1})p(N|\omega^B_2) + q_{t-1}p(I|\omega^R_2)] + (1-\alpha)(\mathcal{P}^R_{t-1} - \epsilon\Delta\mathcal{P}_{t-1}) & \text{if } \Delta P_t(\omega_2, \ell_t) < 0. \end{cases}$$

[19] With systematic polarization, more precisely, $P^B_t(\omega_2, \ell^B_t)$ is the minimum of the expression on the RHS and 1.

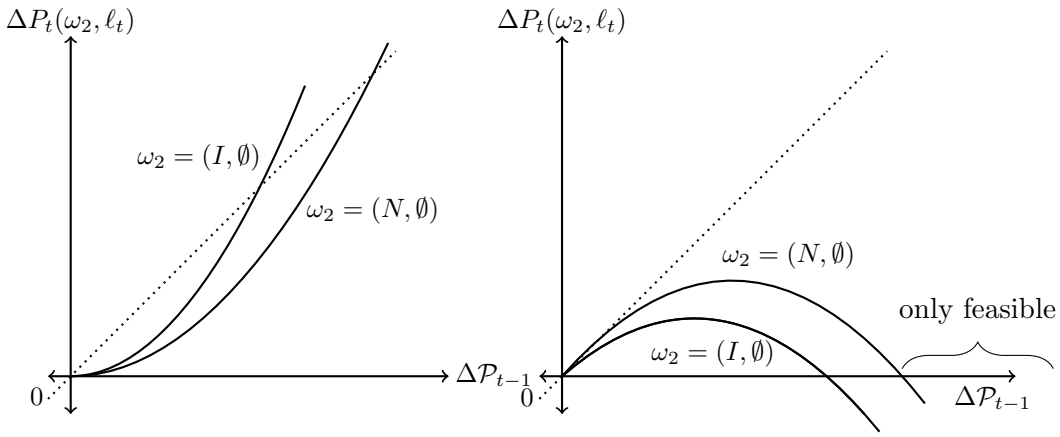

Figure 5: Belief dynamics for $\Delta P_t(\omega_2, \ell_t) > 0$ (left) and for $\Delta P_t(\omega_2, \ell_t) < 0$ (right)

Let $\gamma \equiv 1+2\epsilon > 1$ be the sum of the inertia and the gross amplifying effect in the sense that the first term 1 is from pure inertia and the second term, 2ϵ, comes from the combined amplifying effects of both groups. With this and simple algebra as found in the proof of the following proposition, the same group matching probability q_{t-1} in (4) can be replaced by the term with $\Delta \mathcal{P}_{t-1}$, and the *belief population composition dynamics* can be constructed as follows.

$$\Delta P_t(\omega_2, \ell_t) = \begin{cases} \alpha \frac{1-N_2(\omega)}{A_{t-1}(2-A_{t-1})} \Delta \mathcal{P}_{t-1}^2 + (1-\alpha)\gamma \Delta \mathcal{P}_{t-1} & \text{if } \Delta P_t(\omega_2, \ell_t) > 0, \\ -\alpha \frac{1-N_2(\omega)}{A_{t-1}(2-A_{t-1})} \Delta \mathcal{P}_{t-1}^2 + (1-\alpha)\gamma \Delta \mathcal{P}_{t-1} & \text{if } \Delta P_t(\omega_2, \ell_t) < 0, \end{cases} \tag{25}$$

where we denote the *sum of the two moving probabilities* by $A_{t-1} \equiv \mathcal{P}_{t-1}^B + \mathcal{P}_{t-1}^R$. That is, if $\alpha = 1$, the above formula is identical to the beliefs dynamics only with strategic choices in (23) in the previous section. Note that the more frequent matching with members of the same group as shown in Lemma 1 generates a *strictly convex* shape for the belief dynamics in (25), as illustrated in Figure 5, which will spill over into the real dynamics as well (as can be seen in Figures 6 and 7).[20]

In the belief dynamics, one can find that even if $\Delta \mathcal{P}_{t-1} > 0$, we could have $\Delta P_t(\omega_2, \ell_t) < 0$ (the right panel of Figure 5), as well as $\Delta P_t(\omega_2, \ell_t) > 0$, but if the negative sign arises, it does so only when $\Delta \mathcal{P}_{t-1} > 0$ is sufficiently large, with the horizontal intercept $\frac{(1-\alpha)\gamma}{\alpha} \frac{A_{t-1}(2-A_{t-1})}{1-N_2(\omega_2)}$; that is, 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. By adopting a Markov strategy depending only on $\Delta \mathcal{P}_{t-1}$ we erase such unreasonable, non-robust outcomes. In other words, if $\Delta \mathcal{P}_{t-1} > 0$, the belief dynamics follow only the left panel of Figure 5, whereas if $\Delta \mathcal{P}_{t-1} < 0$, the belief dynamics follow the right panel of Figure 5 (negative side of it). Then, extending $\ell_t^g(\omega_2)$ in (41), a group $g \in \{B, R\}$ player's location decision at the interim period between $t-1$ and t is given by a Markov strategy if it is a function of ω_2 and $\Delta \mathcal{P}_{t-1}$ such that $\ell_t^g : \Omega_2 \times [0,1] \to \{E, W\}$. Then, even with systematic polarization, considering all α, the necessary and sufficient condition for the existence of a binary splitting equilibrium is the same

[20]Note that for Figures 5-7, in the case of $\Delta P_t(\omega_2, \ell_t) < 0$, by symmetry, the negative dimension's graph looks exactly the same as the reversed one in the positive dimension in the case of $\Delta P_t(\omega_2, \ell_t) > 0$.

as the one from Proposition 4.[21] A *Markov binary splitting location equilibrium* ℓ_t is given as: for each $\omega_2 \in \Omega_2^{B+} = \Omega_2^{R-}$, $\omega_2' \in \Omega_2^{B-} = \Omega_2^{R+}$,

$$\begin{cases} \ell_t^g(\omega_2, \Delta\mathcal{P}_{t-1}) = E, \ell_t^g(\omega_2', \Delta\mathcal{P}_{t-1}) = W & \text{if } \Delta\mathcal{P}_{t-1} > 0, \\ \ell_t^g(\omega_2, \Delta\mathcal{P}_{t-1}) = W, \ell_t^g(\omega_2', \Delta\mathcal{P}_{t-1}) = E & \text{if } \Delta\mathcal{P}_{t-1} < 0. \end{cases} \tag{26}$$

The proposition below shows that the above Markov strategy is only robust in the sense that the period t equilibrium belief, $\Delta P_t(\omega_2, \ell_t) > 0$ or $\Delta P_t(\omega_2, \ell_t) < 0$ or both, does not change depending on the specific value of $\Delta\mathcal{P}_{t-1}$; that is, the sign is the same regardless of $\Delta\mathcal{P}_{t-1}$.[22] Further, it makes the dynamics include the limit point $\mathcal{P}_t^B = \mathcal{P}_t^R$ by incorporating the case $P_t^B(\omega_2, \ell_t^B) = P_t^R(\omega_2, \ell_t^R)$.

Proposition 5 *Suppose (A1)-(A5) and $N_2(N, \emptyset) < 1$. Then, for each $\Delta\mathcal{P}_{t-1}$ for $t = 1, 2, 3, ...$,*

(i) Any robust location equilibrium is a Markov location equilibrium.

(ii) $P_t^B(\omega_2, \ell_t^B) = P_t^R(\omega_2, \ell_t^R)$ if and only if $\mathcal{P}_t^B = \mathcal{P}_t^R$ for a Markov location equilibrium.

We incorporate this Markov location equilibrium into the actual dynamics with the two true distributions, F_B and F_R. In other words, we turn the belief dynamics in (25) into real dynamics by replacing $N_2(\omega)$ in (24) by $\mathcal{N}_2$, where we denote $\mathcal{N}_2 \equiv F_B(k_1) + F_R(k_1)$. With systematic polarization, *i.e.*, when $\gamma > 1$, the *real population composition dynamics* are given by $\Delta\mathcal{P}_t = f_{t-1}\big(\Delta\mathcal{P}_{t-1}\big)$ with a function f_{t-1} for $t = 1, 2, ...$ such that

$$f_{t-1}(\Delta\mathcal{P}_{t-1}) \equiv \begin{cases} \alpha\beta_{t-1}\Delta\mathcal{P}_{t-1}^2 + (1-\alpha)\gamma\Delta\mathcal{P}_{t-1} & \text{if } \Delta\mathcal{P}_{t-1} \geq 0, \\ -\alpha\beta_{t-1}\Delta\mathcal{P}_{t-1}^2 + (1-\alpha)\gamma\Delta\mathcal{P}_{t-1} & \text{if } \Delta\mathcal{P}_{t-1} \leq 0, \end{cases} \tag{27}$$

where we denote

$$\beta_{t-1} \equiv \frac{1 - \mathcal{N}_2}{A_{t-1}(2 - A_{t-1})}. \tag{28}$$

In spite of some procedures for a location equilibrium given the uncertainty about types and distributions along with private monitoring, we obtain a simple main dynamical system with a *square function*, in which denote each fixed point of f_{t-1} by x_{t-1}^*, *i.e.*, $x_{t-1}^* = f_{t-1}(x_{t-1}^*)$, and it is derived as

$$x_{t-1}^* = \frac{1 - (1-\alpha)\gamma}{\alpha\beta_{t-1}}. \tag{29}$$

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 of Proposition 5 as a population *flow*. Yet, this system is different from any standard dynamics in that the fixed point of the system keeps moving. As a result, there is no *interior absorbing* state. In addition, the fixed point depends on $\mathcal{N}_2$, that is, the nature of the population distributions, what combination – among the four possible combinations – (F_B, F_R) can take. In particular, the equilibrium existence condition $1 - N_2(N, \emptyset) > 0$ does not necessarily imply

[21] That is, without this condition, for a sufficiently small $1 - \alpha$, no equilibrium exists, based on the same proof from Proposition 4.

[22] We can strengthened it by requiring the horizontal intercept in the right panel of Figure 5 greater than 1. Then, for any $\Delta\mathcal{P}_{t-1} > 0$, it is *always* the case that $\Delta P_t(\omega_2, \ell_t) > 0$, so that $\Delta P_t(\omega_2, \ell_t) < 0$ never arises.

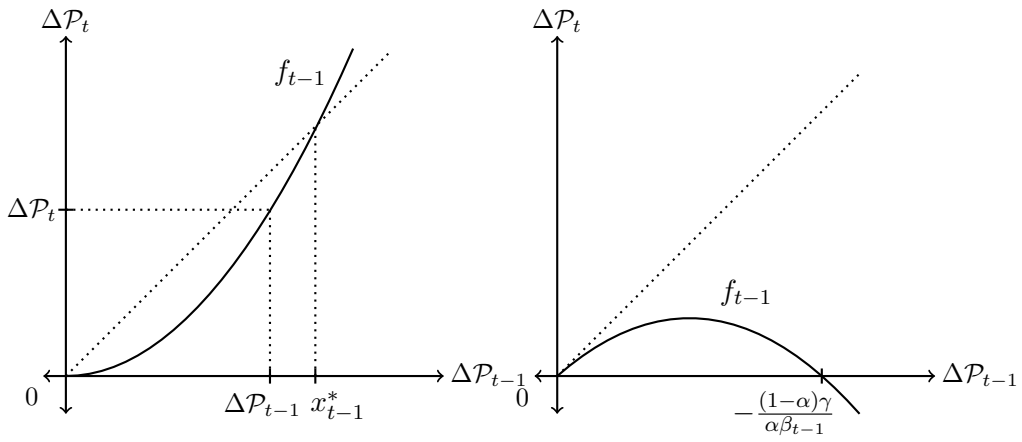

Figure 6: Real dynamics for convergence if $1-\mathcal{N}_2>0$ (left) and $1-\mathcal{N}_2<0$ (right)

$1-\mathcal{N}_2>0$ based on the true distributions: $1-\mathcal{N}_2<0$ can arise if $(F_B,F_R)=(F_Y,F_Y)$ – it is a necessary condition – even when $\Delta\mathcal{P}_{t-1}>0$, $\Delta\mathcal{P}_t<0$ can arise.[23]

As shown in the following lemma, the sum of the two probabilities $\mathcal{P}_t^B$ and $\mathcal{P}_t^R$ evolves as follows:

$$A_t=\begin{cases}\alpha[1-F_B(k_1)+F_R(k_1)]+(1-\alpha)A_{t-1} & \text{if } \Delta\mathcal{P}_{t-1}\geq 0,\\ \alpha[1+F_B(k_1)-F_R(k_1)]+(1-\alpha)A_{t-1} & \text{if } \Delta\mathcal{P}_{t-1}\leq 0.\end{cases} \tag{30}$$

Each fixed point being dependent on β_t and the quadratic functional form of A_t in the denominator of β_t in (28), Lemma 5 shows that if an initial population difference based on A_0 is not in the middle, a sequence of fixed points satisfies monotonicity: x_t^* strictly increases.[24]

Lemma 5 *Suppose (A1)-(A5) and a Markov location equilibrium. A sequence of fixed points satisfies that if* $A_0<1-|F_B(k_1)-F_R(k_1)|$ *or* $A_0>1+|F_B(k_1)-F_R(k_1)|$, $x_t^*>x_{t-1}^*$ *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 *all* initial A_0.

As depicted in Figure 6, whether there is polarization hinges on the relationship between the fixed point for the real dynamics function in period $t-1$ and the state variable $\Delta\mathcal{P}_{t-1}$, the difference in the proportion at E between the two groups in that period. By Lemma 5, 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.

We are now ready to characterize the limiting 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.

[23] If $1-\mathcal{N}_2=0$ in (27), there is no endogenous part in the dynamics; this is a trivial and not-interesting case.

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

Proposition 6 *(Completely mixed state) Suppose (A1)-(A5) and a Markov location equilibrium. Then, for A_0 satisfying Lemma 5, 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 $\lim_{t\geq\tau} \Delta\mathcal{P}_t = 0$.*

Note that the initial condition for A_0 is always satisfied, if we have the same type distribution, $F_B = F_R$. The most interesting case for the completely mixed state is when there is no systematic polarization, $\gamma = 1$. That is, if all location choices are made strategically without any exogenous force, the society converges to a completely mixed state, despite the dynamical system having a square function. To delve into this case further, we examine $\beta_{\tau-1}\Delta\mathcal{P}_{\tau-1}$, and by $A_{t-1} \equiv \mathcal{P}^B_{t-1} + \mathcal{P}^R_{t-1}$ in (25) and β_{t-1} in (28),

$$\beta_{\tau-1}\Delta\mathcal{P}_{\tau-1} = \frac{1-\mathcal{N}_2}{A_{\tau-1}(2-A_{\tau-1})}\Delta\mathcal{P}_{\tau-1} = (1-\mathcal{N}_2)\frac{\mathcal{P}^B_{t-1} - \mathcal{P}^R_{t-1}}{(\mathcal{P}^B_{t-1} + \mathcal{P}^R_{t-1})(2 - \mathcal{P}^B_{t-1} - \mathcal{P}^R_{t-1})},$$

where for any $\mathcal{P}^B_{t-1}, \mathcal{P}^R_{t-1} \in [0,1]$, the fraction in the last line above is always less than or equal to 1. Hence, for each $\tau \geq 1$, we have $\Delta\mathcal{P}_{\tau-1} < \frac{1}{\beta_{\tau-1}} = x^*_{\tau-1}$. Each period's state variable $\Delta\mathcal{P}$, representing the difference in the moving probabilities of the two groups, is always smaller than the corresponding fixed point $x^*_{\tau-1}$ of the system. The intuition is that despite the square function, the difference between two moving probabilities is *generically* lower than the "base" from their sum, $A_{\tau-1}(2 - A_{\tau-1})$.

6.1 Polarization: same type distribution

Suppose the two groups have the same type distribution, $F_B = F_R$. Then, from (30), the sum of the two probabilities $\mathcal{P}^B_t$ and $\mathcal{P}^R_t$ is given as $A_t = \alpha + (1-\alpha)A_{t-1}$.

We can now present our main results for polarization in this case of the same type distributions. Without loss of generality, we consider $\Delta\mathcal{P}_0 > 0$ in what follows.[25] 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, and never be caught by the latter. 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. Precisely, for each size of systematic polarization, there exists a corresponding initial condition 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 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 the polarization – and what is interesting is, of course, the latter.[26]

Proposition 7 *(Polarization: same distribution) Suppose (A1)-(A5) and a Markov location equilibrium. Then, the polarization results are given as follows.*

[25] 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 $1 - \mathcal{P}^B_0 > 1 - \mathcal{P}^R_0$, the same analysis applies, now, in terms of a B player's choice to locate in the *West*.

[26] That is, if $(1-\alpha)\gamma \geq 1$, the exogenous transition *by itself* makes the society converge to the polarization outcome. Thus, a non-trivial case is when $(1-\alpha)\gamma < 1$.

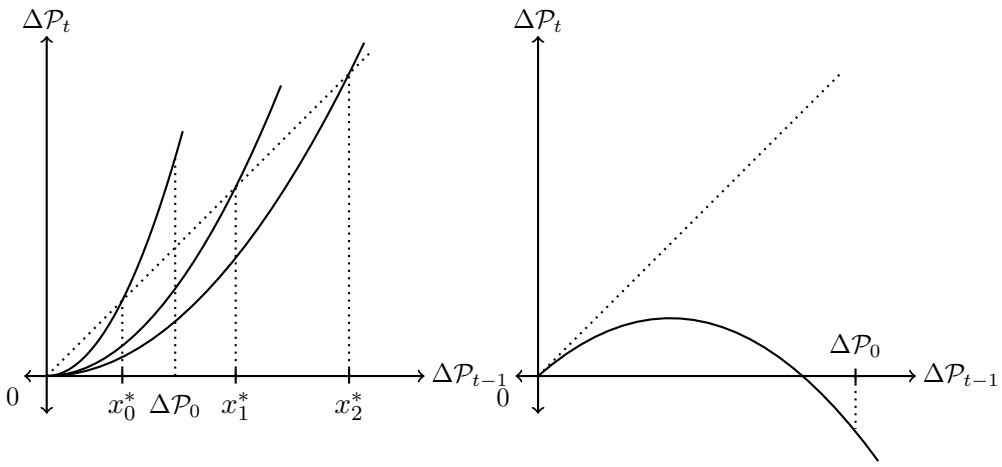

Figure 7: Real dynamics for polarization if $1-\mathcal{N}_2>0$ (left) and $1-\mathcal{N}_2<0$ (right)

(i) If $1-\mathcal{N}_2>0$, then for each $(1-\alpha)\gamma$, there exists a critical value $x_S^\dagger<1$ such that the society converges to a completely polarized state if $\Delta\mathcal{P}_0>x_S^\dagger$.

(ii) If $1-\mathcal{N}_2<0$, then for each $(1-\alpha)\gamma$, there exists $x_S^{\dagger\dagger}<1$ such that the society converges to a completely polarized state if $\Delta\mathcal{P}_0>x_S^{\dagger\dagger}$.

The resulting 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 "bad."[27]

Note that if the amplifying effect γ is relatively small, by (29), the fixed point for each t becomes larger, resulting in a higher critical value. Hence, the smaller 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.

6.2 Polarization: different type distributions

We now make a comparison between the case in the previous section where the distributions are the same, $F_B=F_R$, and 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 it 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

[27]From the right panel of Figure 7, for $\Delta\mathcal{P}_0>0$, if $\Delta\mathcal{P}_1<1$, then $f_1(x)=-\alpha\beta_1x^2+(1-\alpha)\gamma x$ from (27) and so on. Like the no-catching up condition in the left panel of Figure 7, we find the no-catching up condition for the oscillation case.

of Proposition 7, oscillation, only arises with $(F_B, F_R) = (F_Y, F_Y)$ – the comparison with different distributions is not possible, so the following result is written only for $1 - \mathcal{N}_2 > 0$.

Proposition 8 *(Polarization: different distributions) Suppose (A1)-(A5) and a Markov location equilibrium. Then, for $1 - \mathcal{N}_2 > 0$, the polarization results are given as follows.*

(i) If $A_0 > 1$, the critical value $x_A^\dagger$ for the different distributions $(F_B, F_R) = (F_X, F_Y)$ is lower (resp. higher) than the critical value $x_S^\dagger$ 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 $x_A^\dagger$ for the different distributions $(F_B, F_R) = (F_Y, F_X)$ is lower (resp. higher) than the critical value $x_S^\dagger$ 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 among the total B members, B group's better distribution facilitates polarization more, compared with 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 among the total R members. Likewise, this together with R group's better distribution facilitates polarization.

7 Discussion

We can expand upon our results in four dimensions. First, a unique equilibrium can be guaranteed if (A3) is strengthened such that for each $\theta' > \theta$ in Γ, $d(\theta') - d(\theta) \geq \pi[F_X(\theta') - F_X(\theta)] + (1 - \pi)[F_Y(\theta') - F_Y(\theta)], \pi = \pi_2(I), \pi_2(N)$. Now, the relationship holds not only for the biased belief but for $\pi_2(I), \pi_2(N)$. 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 m is the same as the one from (A5).

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.[28] 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 for revelation of 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 $f_{t-1}(x, \widetilde{\alpha}_{t-1})$, not $f_{t-1}(x)$ in (27).

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 (27)

[28] 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."

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.[29]

Finally, we have not yet discussed what happens after there is a complete polarization. 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 (7) changes to:

$$\begin{aligned} d(k_t^S(\omega_2^S)) = {} & \pi(\omega_2^S)[(1 - F_X(k_1))F_X(k_t^S(I)) + F_X(k_1)F_X(k_t^S(N))] \\ & + (1 - \pi(\omega_2^S))[(1 - F_Y(k_1))F_Y(k_t^S(I)) + F_Y(k_1)F_Y(k_t^S(N))], \end{aligned}$$

which provides the expected payoff for history I and N from meeting a member of the same group by still remaining in the same location. On the other hand, moving to the other location yields the expected payoff (3) from meeting with the other group member. Since the comparison between the two mappings in Figure 4 still holds in the limit, that is, $\lim q_{t-1} = 1$, we have $k(I) < k_1 < k(N)$. 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 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 x_{t-1}^* for the polarization in Proposition 7, a *necessary* condition is that $x_{t-1}^* = \frac{1-(1-\alpha)\gamma}{\alpha\beta_{t-1}} < 1$, which can be rewritten as $\frac{1-(1-\alpha)\gamma}{\alpha(1-\mathcal{N}_2)} < A_{t-1}(2 - A_{t-1}) \leq 1$. This in turn implies that for $\Delta\mathcal{P}_{t-1} = 1$ and $A_{t-1} = 1$, $\alpha(1-\mathcal{N}_2) + (1-\alpha)\gamma > 1$ in (27), so the minimum of $\alpha(1-\mathcal{N}_2) + (1-\alpha)\gamma$ and 1 is still 1: The endogenous move $\alpha(1-\mathcal{N}_2)$ combined with the exogenous force maintains the complete polarization. In other words, despite the endogenous decisions, the amplifying effects make no young player explore the other group.

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. Agents live two periods and initially live in one of two locations, the one in which they are born. They interact only with members of their own generation in play of an investment stage game. Based

[29] To elaborate, this is true for a *non-trivial* case $\delta(1-\alpha)\gamma < 1$ so 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.

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 obtains even if the two groups have the *same* type distribution. We further show that convergence to such a polarized outcome depends on the size of the amplifying force; if this amplifying force is sufficiently small, then the long-term outcome of the system is 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

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

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

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

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

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

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

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

Proof of Proposition 1. We start with deriving the probability of a matched partner choosing No Invest $\Pr(N|\omega_2^S, \mathbf{k}_t^S)$ in (5) precisely. The probability is given by

$$\begin{aligned}\Pr(N|\omega_2^S, \mathbf{k}_t^S) &= \Pr(N|\omega_2^S, \mathbf{k}_t^S) = \Pr(N, X^S|\omega_2^S, \mathbf{k}_t^S) + \Pr(N, Y^S|\omega_2^S, \mathbf{k}_t^S) \\ &= \Pr(X^S|\omega_2^S, \mathbf{k}_t^S)\Pr(N|X^S, \omega_2^S, \mathbf{k}_t^S) + \Pr(Y^S|\omega_2^S, \mathbf{k}_t^S)\Pr(N|Y^S, \omega_2^S, \mathbf{k}_t^S).\end{aligned} \tag{31}$$

For each $r \in \{X, Y\}$, $\Pr(r^S|\omega_2^S, \mathbf{k}_t^S)$ is the probability that the same group's distribution is F_r and $\Pr(N|r^S, \omega_2^S, \mathbf{k}_t^S)$ is the probability that the partner chooses No Invest, conditional on F_r. In particular, the latter probability depends on previous observations of the partner ω_2^S. We denote $X_t^S(\mathbf{k}_t^S) \equiv \Pr(N|X^S, \omega_2^S, \mathbf{k}_t^S)$ and $Y_t^S(\mathbf{k}_t^S) \equiv \Pr(N|Y^S, \omega_2^S, \mathbf{k}_t^S)$ and derive them as (6). Since for $\omega_2^S = I, N \in \Omega_2^S$, $\Pr(X^S|\omega_2^S, \mathbf{k}_t^S) = \pi_2(\omega_2^S)$ in (2), and no previous match with B player yields no updating on B group, $\pi_2(\emptyset) = \frac{1}{2}$. Together, we have

$$\Pr(N|\omega_2^S, \mathbf{k}_t^S) = \pi_2(\omega_2^S)X_t^S(\mathbf{k}_t^S) + (1 - \pi_2(\omega_2^S))Y_t^S(\mathbf{k}_t^S).$$

It suffices to show that for any pair of two histories $\widehat{\omega}_2^S, \omega_2^S \in \Omega_2^S$ with $\pi_2(\widehat{\omega}_2^S) > \pi_2(\omega_2^S)$, we have $k_2(\widehat{\omega}_2^S) < k_2(\omega_2^S)$. Suppose, on the contrary, that $k_2(\widehat{\omega}_2^S) \geq k_2(\omega_2^S)$. From (7), any pair of two histories $\widehat{\omega}_2^S, \omega_2^S \in \Omega_2^S$ yields a difference in two thresholds such that

$$d(k_t^S(\widehat{\omega}_2^S)) - d(k_t^S(\omega_2^S)) = \left(\pi_2(\widehat{\omega}_2^S) - \pi_2(\omega_2^S)\right)\left[X_t^S(\mathbf{k}_t^S) - Y_t^S(\mathbf{k}_t^S)\right]. \tag{32}$$

We divide the proof into two cases.

Case 1. $k_t^S(\widehat{\omega}_2^S) = k_t^S(\omega_2^S)$. The difference in (32) results in $X_t^S(\mathbf{k}_t^S) = Y_t^S(\mathbf{k}_t^S)$, which in turn implies that $k_t = k_t^S(\omega_2^S)$ for all ω_2^S. Then, it follows from (6) that $X_t^S(\mathbf{k}_t^S) - Y_t^S(\mathbf{k}_t^S) = F_X(k_t) - F_Y(k_t)$, so

$$\begin{aligned}0 = d(k_t^S(\widehat{\omega}_2^S)) - d(k_t^S(\omega_2^S)) &= \left(\pi_2(\widehat{\omega}_2^S) - \pi_2(\omega_2^S)\right)\left[X_t^S(\mathbf{k}_t^S) - Y_t^S(\mathbf{k}_t^S)\right] \\ &= \left(\pi_2(\widehat{\omega}_2^S) - \pi_2(\omega_2^S)\right)[F_X(k_t) - F_Y(k_t)] < 0,\end{aligned}$$

which is a contradiction.

Case 2. $k_t^S(\widehat{\omega}_2^S) > k_t^S(\omega_2^S)$. The difference in (32) and $\pi_2(\widehat{\omega}_2^S) > \pi_2(\omega_2^S)$ 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

$$d\left(k_t^{\max}\right) - d\left(k_t^{\max}\right) = \left(\pi_2^{\max} - \pi_2^{\min}\right)\left[X_t^S(\mathbf{k}_t^S) - Y_t^S(\mathbf{k}_t^S)\right].$$

From $X_t^S(\mathbf{k}_t^S) > Y_t^S(\mathbf{k}_t^S)$ and $0 < \pi_2^{\max} - \pi_2^{\min} < 1$, for each $r \in \{X, Y\}$, we have

$$d\left(k_t^{\max}\right) - d\left(k_t^{\max}\right) = \left(\pi_2^{\max} - \pi_2^{\min}\right)\left[X_t^S(\mathbf{k}_t^S) - Y_t^S(\mathbf{k}_t^S)\right] < \left[X_t^S(\mathbf{k}_t^S) - Y_t^S(\mathbf{k}_t^S)\right] < F_r(k_t^{\max}) - F_r(k_t^{\min}),$$

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 Lemma 2. We start by deriving the probability of a matched partner choosing No Invest $\Pr(N|\omega_2^A, \mathbf{k}_t^A)$ in (9) precisely. We put a group-specific superscript $\pi(I^g)$ or $\pi(N^g)$ for $g \in \{B, R\}$ if doing so helps us keep track of which group player chooses a corresponding action in what follows. Without loss of generality, let us examine the problem from the B member's perspective to investigate the No Invest probability $\Pr(N|\omega_2^A, \mathbf{k}_t^A)$, which is given as

$$\begin{aligned}\Pr(N|\omega_2^A, \mathbf{k}_t^A) = &\Pr(N, (X^B, X^R)|\omega_2^A, \mathbf{k}_t^A) + \Pr(N, (X^B, Y^R)|\omega_2^A, \mathbf{k}_t^A) \\ &+ \Pr(N, (Y^B, X^R)|\omega_2^A, \mathbf{k}_t^A) + \Pr(N, (Y^B, Y^R)|\omega_2^A, \mathbf{k}_t^A) \\ = &\Pr(X^R|\omega_2^A, \mathbf{k}_t^A)\left[\Pr(N, X^B|X^R, \omega_2^A, \mathbf{k}_t^A) + \Pr(N, Y^B|X^R, \omega_2^A, \mathbf{k}_t^A)\right] \\ &+ \Pr(Y^R|\omega_2^A, \mathbf{k}_t^A)\left[\Pr(N, X^B|Y^R, \omega_2^A, \mathbf{k}_t^A) + \Pr(N, Y^B|Y^R, \omega_2^A, \mathbf{k}_t^A)\right].\end{aligned} \tag{33}$$

Furthermore, denote

$$\begin{aligned}X_t^A(\mathbf{k}_t^A, \omega_2^B) &\equiv \Pr(N, X^B|X^R, \omega_2^A, \mathbf{k}_t^A) + \Pr(N, Y^B|X^R, \omega_2^A, \mathbf{k}_t^A), \\ Y_t^A(\mathbf{k}_t^A, \omega_2^B) &\equiv \Pr(N, X^B|Y^R, \omega_2^A, \mathbf{k}_t^A) + \Pr(N, Y^B|Y^R, \omega_2^A, \mathbf{k}_t^A).\end{aligned}$$

We now derive $X_t^A(\mathbf{k}_t^A, \omega_2^B)$ and $Y_t^A(\mathbf{k}_t^A, \omega_2^B)$ with $p_2(I|\omega_2^B)$ and $p_2(N|\omega_2^B)$ such that

$$\begin{aligned}X_t^A(\mathbf{k}_t^A, \omega_2^B) = &\, q_{t-1}(1 - F_X(k_1))F_X(k_t^A(\emptyset|I)) + q_{t-1}F_X(k_1)F_X(k_t^A(\emptyset|N)) \\ &+ (1 - q_{t-1})p_2(I|\omega_2^B)F_X(k_t^A(I|\emptyset)) + (1 - q_{t-1})p_2(N|\omega_2^B)F_X(k_t^A(N|\emptyset)), \\ Y_t^A(\mathbf{k}_t^A, \omega_2^B) = &\, q_{t-1}(1 - F_Y(k_1))F_Y(k_t^A(\emptyset|I)) + q_{t-1}F_Y(k_1)F_Y(k_t^A(\emptyset|N)) \\ &+ (1 - q_{t-1})p_2(I|\omega_2^B)F_Y(k_t^A(I|\emptyset)) + (1 - q_{t-1})p_2(N|\omega_2^B)F_Y(k_t^A(N|\emptyset)).\end{aligned} \tag{34}$$

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 $(1 - F_r(k_1))$, he observed Invest from that partner, which with a corresponding threshold $k_t^A(\emptyset|I)$ yields the first term $q_{t-1}(1-F_X(k_1))F_X(k_t^A(\emptyset|I))$. In addition, with probability $1-q_{t-1}$, the matched R player met a B player previously when young, and with probability $p_2(I|\omega_2^B)$, he observed Invest from that partner, which with a corresponding threshold $k_t^A(I|\emptyset)$ yields the third term $(1 - q_{t-1})p_2(I|\omega_2^B)F_X(k_t^A(I|\emptyset))$. The other two terms can be readily derived accordingly. Since for $\omega_2^R = I, N$ from $\omega_2^A \in \Omega_2^A$, $\Pr(X^R|\omega_2^R, \mathbf{k}_t^S) = \pi_2(\omega_2^R)$ in (2) together with $\pi_2(\emptyset) = \frac{1}{2}$, we have

$$\Pr(N|\omega_2^A, \mathbf{k}_t^A) = \pi_2(\omega_2^R)X_t^A(\mathbf{k}_t^A, \omega_2^B) + (1 - \pi_2(\omega_2^R))Y_t^A(\mathbf{k}_t^A, \omega_2^B).$$

(i), (ii) Take the difference between $k_t^A(I|\emptyset)$ and $k_t^A(\emptyset|I)$ such that

$$\begin{aligned}&d(k_t^A(I|\emptyset)) - d(k_t^A(\emptyset|I)) \\ &= \pi_2(I^R)X_t^A(\mathbf{k}_t^A, \emptyset^B) + (1 - \pi_2(I^R))Y_t^A(\mathbf{k}_t^A, \emptyset^B) - [\pi_2(\emptyset^R)X_t^A(\mathbf{k}_t^A, I^B) + (1 - \pi_2(\emptyset^R))Y_t^A(\mathbf{k}_t^A, I^B)],\end{aligned}$$

where the RHS can be expanded as follows:

$$\pi_2(I^R)\left[\begin{array}{l} q_{t-1}(1-F_X(k_1))F_X(k_t^A(\emptyset|I))+q_{t-1}F_X(k_1)F_X(k_t^A(\emptyset|N)) \\ +(1-q_{t-1})\big[\pi_2(\emptyset^B)(1-F_X(k_1))+(1-\pi_2(\emptyset^B))(1-F_Y(k_1))\big]F_X(k_t^A(I|\emptyset)) \\ +(1-q_{t-1})\big[\pi_2(\emptyset^B)F_X(k_1)+(1-\pi_2(\emptyset^B))F_Y(k_1)\big]F_X(k_t^A(N|\emptyset)) \end{array}\right] \tag{35}$$

$$+(1-\pi_2(I^R))\left[\begin{array}{l} q_{t-1}(1-F_Y(k_1))F_Y(k_t^A(\emptyset|I))+q_{t-1}F_Y(k_1)F_Y(k_t^A(\emptyset|N)) \\ +(1-q_{t-1})\big[\pi_2(\emptyset^B)(1-F_X(k_1))+(1-\pi_2(\emptyset^B))(1-F_Y(k_1))\big]F_Y(k_t^A(I|\emptyset)) \\ +(1-q_{t-1})\big[\pi_2(\emptyset^B)F_X(k_1)+(1-\pi_2(\emptyset^B))F_Y(k_1)\big]F_Y(k_t^A(N|\emptyset)) \end{array}\right]$$

$$-\pi_2(\emptyset^R)\left[\begin{array}{l} q_{t-1}(1-F_X(k_1))F_X(k_t^A(\emptyset|I))+q_{t-1}F_X(k_1)F_X(k_t^A(\emptyset|N)) \\ +(1-q_{t-1})\big[\pi_2(I^B)(1-F_X(k_1))+(1-\pi_2(I^B))(1-F_Y(k_1))\big]F_X(k_t^A(I|\emptyset)) \\ +(1-q_{t-1})\big[\pi_2(I^B)F_X(k_1)+(1-\pi_2(I^B))F_Y(k_1)\big]F_X(k_t^A(N|\emptyset)) \end{array}\right] \tag{36}$$

$$-(1-\pi_2(\emptyset^R))\left[\begin{array}{l} q_{t-1}(1-F_Y(k_1))F_Y(k_t^A(\emptyset|I))+q_{t-1}F_Y(k_1)F_Y(k_t^A(\emptyset|N)) \\ +(1-q_{t-1})\big[\pi_2(I^B)(1-F_X(k_1))+(1-\pi_2(I^B))(1-F_Y(k_1))\big]F_Y(k_t^A(I|\emptyset)) \\ +(1-q_{t-1})\big[\pi_2(I^B)F_X(k_1)+(1-\pi_2(I^B))F_Y(k_1)\big]F_Y(k_t^A(N|\emptyset)) \end{array}\right].$$

Since $\pi_2(I^R)\pi_2(\emptyset^B)=\pi_2(I^B)\pi_2(\emptyset^R)$, the third term of the bracket in (35) and the third term of the bracket in (36) can be simplified such that

$$\begin{aligned} &\pi_2(I^R)(1-q_{t-1})\big[\pi_2(\emptyset^B)(1-F_X(k_1))+(1-\pi_2(\emptyset^B))(1-F_Y(k_1))\big]F_X(k_t^A(I|\emptyset)) \\ &-\pi_2(\emptyset^R)(1-q_{t-1})\big[\pi_2(I^B)(1-F_X(k_1))+(1-\pi_2(I^B))(1-F_Y(k_1))\big]F_X(k_t^A(I|\emptyset)) \\ &=\pi_2(I^R)(1-q_{t-1})(1-F_Y(k_1))F_X(k_t^A(I|\emptyset))-\pi_2(\emptyset^R)(1-q_{t-1})(1-F_Y(k_1))F_X(k_t^A(I|\emptyset)). \end{aligned} \tag{37}$$

The same type of deletion can be applied to the fourth term of the bracket in (35) and the fourth term of the bracket in (36). Furthermore, considering the bracket following $(1-\pi_2(I^R))$ and the bracket following $(1-\pi_2(\emptyset^R))$, the whole formula can be rewritten as

$$\begin{aligned} &\pi_2(I^R)\left[\begin{array}{l} q_{t-1}(1-F_X(k_1))F_X(k_t^A(\emptyset|I))+q_{t-1}F_X(k_1)F_X(k_t^A(\emptyset|N)) \\ +(1-q_{t-1})(1-F_Y(k_1))F_X(k_t^A(I|\emptyset))+(1-q_{t-1})F_Y(k_1)F_X(k_t^A(N|\emptyset)) \end{array}\right] \\ &+(1-\pi_2(I^R))[q_{t-1}(1-F_Y(k_1))F_Y(k_t^A(\emptyset|I))+q_{t-1}F_Y(k_1)F_Y(k_t^A(\emptyset|N))] \\ &-\pi_2(\emptyset^R)\left[\begin{array}{l} q_{t-1}(1-F_X(k_1))F_X(k_t^A(\emptyset|I))+q_{t-1}F_X(k_1)F_X(k_t^A(\emptyset|N)) \\ +(1-q_{t-1})(1-F_Y(k_1))F_X(k_t^A(I|\emptyset))+(1-q_{t-1})F_Y(k_1)F_X(k_t^A(N|\emptyset)) \end{array}\right] \\ &-(1-\pi_2(\emptyset^R))[q_{t-1}(1-F_Y(k_1))F_Y(k_t^A(\emptyset|I))+q_{t-1}F_Y(k_1)F_Y(k_t^A(\emptyset|N))] \\ &+(1-q_{t-1})(\pi_2(\emptyset^R)-(\pi_2(I^R))(1-F_X(k_1))F_Y(k_t^A(I|\emptyset)) \\ &+(1-q_{t-1})(\pi_2(\emptyset^R)-(\pi_2(I^R))F_X(k_1)F_Y(k_t^A(N|\emptyset)), \end{aligned}$$

where the last two terms are rewritten from terms with $(1-q_{t-1})$ in the bracket following $(1-\pi_2(I^R))$ and in the bracket following $(1-\pi_2(\emptyset^R))$, using $(1-\pi_2(I^R))(1-\pi_2(\emptyset^B))=(1-\pi_2(I^B))(1-\pi_2(\emptyset^R))$ and a procedure similar to (37). Additionally, we simplify terms for $F_Y(k_t^A(I|\emptyset))$ and those for $F_Y(k_t^A(N|\emptyset))$ such that

$$\begin{aligned} &(1-\pi_2(I^R))(1-q_{t-1})\big[\pi_2(\emptyset^B)(1-F_X(k_1))+(1-\pi_2(\emptyset^B))(1-F_Y(k_1))\big]F_Y(k_t^A(I|\emptyset)) \\ &-(1-\pi_2(\emptyset^R))(1-q_{t-1})\big[\pi_2(I^B)(1-F_X(k_1))+(1-\pi_2(I^B))(1-F_Y(k_1))\big]F_Y(k_t^A(I|\emptyset)) \\ &=(1-q_{t-1})(\pi_2(\emptyset^R)-(\pi_2(I^R))(1-F_X(k_1))F_Y(k_t^A(I|\emptyset)), \end{aligned}$$

and also

$$\begin{aligned}
&(1-\pi_2(I^R))(1-q_{t-1})\big[\pi_2(\emptyset^B)F_X(k_1)+(1-\pi_2(\emptyset^B))F_Y(k_1)\big]F_Y(k_t^A(N|\emptyset))\\
&-(1-\pi_2(\emptyset^R))(1-q_{t-1})\big[\pi_2(I^B)F_X(k_1)+(1-\pi_2(I^B))F_Y(k_1)\big]F_Y(k_t^A(N|\emptyset))\\
&=(1-q_{t-1})(\pi_2(\emptyset^R)-(\pi_2(I^R))F_X(k_1)F_Y(k_t^A(N|\emptyset)).
\end{aligned}$$

Then, the above is rewritten as

$$\begin{aligned}
&\pi_2(I^R)X_t^A(\mathbf{k}_t^A,\emptyset^B)+(1-\pi_2(I^R))Y_t^A(\mathbf{k}_t^A,\emptyset^B)-[\pi_2(\emptyset^R)X_t^A(\mathbf{k}_t^A,I^B)+(1-\pi_2(\emptyset^R))Y_t^A(\mathbf{k}_t^A,I^B)]\\
&=\pi_2(I^R)\left[\begin{array}{l} q_{t-1}(1-F_X(k_1))F_X(k_t^A(\emptyset|I))+q_{t-1}F_X(k_1)F_X(k_t^A(\emptyset|N))\\ +(1-q_{t-1})(1-F_Y(k_1))F_X(k_t^A(I|\emptyset))+(1-q_{t-1})F_Y(k_1)F_X(k_t^A(N|\emptyset))\end{array}\right]\\
&\quad-\pi_2(\emptyset^R)\left[\begin{array}{l} q_{t-1}(1-F_X(k_1))F_X(k_t^A(\emptyset|I))+q_{t-1}F_X(k_1)F_X(k_t^A(\emptyset|N))\\ +(1-q_{t-1})(1-F_Y(k_1))F_X(k_t^A(I|\emptyset))+(1-q_{t-1})F_Y(k_1)F_X(k_t^A(N|\emptyset))\end{array}\right]\\
&\quad-q_{t-1}(\pi_2(I^R)-\pi_2(\emptyset^R))[(1-F_Y(k_1))F_Y(k_t^A(\emptyset|I))+F_Y(k_1)F_Y(k_t^A(\emptyset|N))]\\
&\quad-(1-q_{t-1})(\pi_2(I^R)-(\pi_2(\emptyset^R))[(1-F_X(k_1))F_Y(k_t^A(I|\emptyset))+F_X(k_1)F_Y(k_t^A(N|\emptyset))],
\end{aligned}$$

which is

$$\begin{aligned}
&q_{t-1}(\pi_2(I^R)-\pi_2(\emptyset^R))[(1-F_X(k_1))F_X(k_t^A(\emptyset|I))+F_X(k_1)F_X(k_t^A(\emptyset|N))]\\
&+(1-q_{t-1})(\pi_2(I^R)-(\pi_2(\emptyset^R))[(1-F_Y(k_1))F_X(k_t^A(I|\emptyset))+F_Y(k_1)F_X(k_t^A(N|\emptyset))]\\
&-q_{t-1}(\pi_2(I^R)-\pi_2(\emptyset^R))[(1-F_Y(k_1))F_Y(k_t^A(\emptyset|I))+F_Y(k_1)F_Y(k_t^A(\emptyset|N))]\\
&-(1-q_{t-1})(\pi_2(I^R)-(\pi_2(\emptyset^R))[(1-F_X(k_1))F_Y(k_t^A(I|\emptyset))+F_X(k_1)F_Y(k_t^A(N|\emptyset))]\\
&<q_{t-1}(\pi_2(I^R)-\pi_2(\emptyset^R))(F_Y(k_1)-F_X(k_1))\left[F_r(k_t^A(\emptyset|I))-F_r(k_t^A(\emptyset|N))\right]\text{ for all } r\in\{X,Y\},
\end{aligned}$$

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

$$\begin{aligned}
&d(k_t^A(\emptyset|I))-d(k_t^A(\emptyset|N))\\
&=\pi_2(\emptyset^R)X_t^A(\mathbf{k}_t^A,I^B)+(1-\pi_2(\emptyset^R))Y_t^A(\mathbf{k}_t^A,I^B)-[\pi_2(\emptyset^R)X_t^A(\mathbf{k}_t^A,N^B)+(1-\pi_2(\emptyset^R))Y_t^A(\mathbf{k}_t^A,N^B)]\\
&=\pi_2(\emptyset^R)(1-q_{t-1})\left[F_X(k_t^A(I|\emptyset))-F_X(k_t^A(N|\emptyset))\right]\big(\pi_2(I^B)-\pi_2(N^B)\big)[F_Y(k_1)-F_X(k_1)]\\
&\quad+(1-\pi_2(\emptyset^R))(1-q_{t-1})\left[F_Y(k_t^A(I|\emptyset))-F_Y(k_t^A(N|\emptyset))\right]\big(\pi_2(I^B)-\pi_2(N^B)\big)[F_Y(k_1)-F_X(k_1)]\\
&=\frac{1}{2}(1-q_{t-1})\big(\pi_2(I^B)-\pi_2(N^B)\big)[F_Y(k_1)-F_X(k_1)]\left[\begin{array}{l} F_X(k_t^A(I|\emptyset))-F_X(k_t^A(N|\emptyset))\\ +F_Y(k_t^A(I|\emptyset))-F_Y(k_t^A(N|\emptyset))\end{array}\right],
\end{aligned}$$

which establishes (iii). ■

Proof of Proposition 2. Consider the difference between $k_t^A(I|\emptyset)$ and $k_t^A(N|\emptyset)$ such that

$$d(k_t^A(I|\emptyset))-d(k_t^A(N|\emptyset))=\big(\pi_2(I^R)-\pi_2(N^R)\big)[X_t^A(\mathbf{k}_t^A,\emptyset^B)-Y_t^A(\mathbf{k}_t^A,\emptyset^B)].$$

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 2 (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 $X_t^A(\mathbf{k}_t^A, \emptyset^B)$, taking those two higher values, and for $Y_t^A(\mathbf{k}_t^A, \emptyset^B)$, 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

$$
\begin{aligned}
&\left[\begin{array}{l} q_{t-1}(1-F_X(k_1))F_X(k_t^A(\emptyset|I)) + q_{t-1}F_X(k_1)F_X(k_t^A(\emptyset|N)) \\ +(1-q_{t-1})\big[\pi_2(\emptyset^B)(1-F_X(k_1)) + (1-\pi_2(\emptyset^B))(1-F_Y(k_1))\big]F_X(k_t^A(I|\emptyset)) \\ +(1-q_{t-1})\big[\pi_2(\emptyset^B)F_X(k_1) + (1-\pi_2(\emptyset^B))F_Y(k_1)\big]F_X(k_t^A(N|\emptyset)) \end{array}\right] \\
&- \left[\begin{array}{l} q_{t-1}(1-F_Y(k_1))F_Y(k_t^A(\emptyset|I)) + q_{t-1}F_Y(k_1)F_Y(k_t^A(\emptyset|N)) \\ +(1-q_{t-1})\big[\pi_2(\emptyset^B)(1-F_X(k_1)) + (1-\pi_2(\emptyset^B))(1-F_Y(k_1))\big]F_Y(k_t^A(I|\emptyset)) \\ +(1-q_{t-1})\big[\pi_2(\emptyset^B)F_X(k_1) + (1-\pi_2(\emptyset^B))F_Y(k_1)\big]F_Y(k_t^A(N|\emptyset)) \end{array}\right] \\
&\leq q_{t-1}F_X(k_t^A(\emptyset|I)) + (1-q_{t-1})F_X(k_t^A(I|\emptyset)) - [q_{t-1}F_Y(k_t^A(\emptyset|N)) + (1-q_{t-1})F_Y(k_t^A(N|\emptyset))] \\
&< q_{t-1}F_r(k_t^A(\emptyset|I)) + (1-q_{t-1})F_r(k_t^A(I|\emptyset)) - [q_{t-1}F_r(k_t^A(\emptyset|N)) + (1-q_{t-1})F_r(k_t^A(N|\emptyset))] \text{ for all } r \in \{X, Y\},
\end{aligned}
$$

where the last inequality follows from the FOSD between F_X and F_Y. Hence, with $0 < \pi_2(I^R) - \pi_2(N^R) < 1$,

$$
\begin{aligned}
&d(k_t^A(I|\emptyset)) - d(k_t^A(N|\emptyset)) \\
&< q_{t-1}F_r(k_t^A(\emptyset|I)) + (1-q_{t-1})F_r(k_t^A(I|\emptyset)) - [q_{t-1}F_r(k_t^A(\emptyset|N)) + (1-q_{t-1})F_r(k_t^A(N|\emptyset))].
\end{aligned} \tag{38}
$$

First, if $k_t^A(I|\emptyset)) = k_t^A(N|\emptyset)$, Lemma 2 (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 2 (iii),

$$
\begin{aligned}
&d(k_t^A(\emptyset|I)) - d(k_t^A(\emptyset|N)) \\
&= \frac{1}{2}(1-q_{t-1})\big(\pi_2(I^B) - \pi_2(N^B)\big)[F_Y(k_1) - F_X(k_1)] \left[\begin{array}{l} F_X(k_t^A(I|\emptyset)) - F_X(k_t^A(N|\emptyset)) \\ +F_Y(k_t^A(I|\emptyset)) - F_Y(k_t^A(N|\emptyset)) \end{array}\right] \\
&< d(k_t^A(I|\emptyset)) - d(k_t^A(N|\emptyset)),
\end{aligned}
$$

where the last inequality follows from (A3) and $0 < (1-q_{t-1})\big(\pi_2(I^B) - \pi_2(N^B)\big)[F_Y(k_1) - F_X(k_1)] < 1$. 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 3. From the homogeneous match in (5), in equilibrium, we have

$$
U_t^S(\theta, \omega_2^S, \mathbf{k}_t^S) \equiv d(\theta) - [\pi_2(\omega_2^S)X_t^S(\mathbf{k}_t^S) + (1-\pi_2(\omega_2^S))Y_t^S(\mathbf{k}_t^S)] = d(\theta) - d(k_t^S(\omega_2^S)),
$$

and from the heterogeneous match in (9), in equilibrium, we have

$$
U_t^A(\theta, \omega_2^A, \mathbf{k}_t^A) \equiv d(\theta) - [\pi_2(\omega_2^R)X_t^A(\mathbf{k}_t^A, \pi_2(\omega_2^B)) + (1-\pi_2(\omega_2^R))Y_t^A(\mathbf{k}_t^A, \pi_2(\omega_2^B))] = d(\theta) - d(k_t^A(\omega_2^A)).
$$

The difference yields the result. Note that

$$
\frac{P_t^B(\omega_2)}{P_t^B(\omega_2) + P_t^R(\omega_2)} - \frac{1 - P_t^B(\omega_2)}{2 - P_t^B(\omega_2) - P_t^R(\omega_2)} = \frac{P_t^B(\omega_2) - P_t^R(\omega_2)}{\big(P_t^B(\omega_2) + P_t^R(\omega_2)\big)\big(2 - P_t^B(\omega_2) - P_t^R(\omega_2)\big)}
$$

and

$$\frac{P_t^R(\omega_2)}{P_t^B(\omega_2)+P_t^R(\omega_2)}-\frac{1-P_t^R(\omega_2)}{2-P_t^B(\omega_2)-P_t^R(\omega_2)}=\frac{P_t^R(\omega_2)-P_t^B(\omega_2)}{\left(P_t^B(\omega_2)+P_t^R(\omega_2)\right)\left(2-P_t^B(\omega_2)-P_t^R(\omega_2)\right)}.$$

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

$$\frac{P_t^B(\omega_2)-P_t^R(\omega_2)}{\left(P_t^B(\omega_2)+P_t^R(\omega_2)\right)\left(2-P_t^B(\omega_2)-P_t^R(\omega_2)\right)}\left[U_t^S(\theta,\omega_2^S,\mathbf{k}_t^S)-U_t^A(\theta,\omega_2^A,\mathbf{k}_t^A)\right].$$

The result follows. ■

Proof of Proposition 3. *Part 1.* We first show (i). With (18) and (19), we define

$$\begin{aligned}\widehat{X}_t(\widehat{\mathbf{k}}_t^\lambda,\omega_2^B,\lambda)\equiv\ & q_{t-1}(1-F_X(k_1))F_X(\widehat{k}_t^\lambda(\emptyset|I))+q_{t-1}F_X(k_1)F_X(\widehat{k}_t^\lambda(\emptyset|N))\\ &+(1-q_{t-1})p_2^X(I|\omega_2^B,\lambda)F_X(\widehat{k}_t^\lambda(I|\emptyset))+(1-q_{t-1})p_2^X(N|\omega_2^B,\lambda)F_X(\widehat{k}_t^\lambda(N|\emptyset)),\\ \widehat{Y}_t(\widehat{\mathbf{k}}_t^\lambda,\omega_2^B,\lambda)\equiv\ & q_{t-1}(1-F_Y(k_1))F_Y(\widehat{k}_t^\lambda(\emptyset|I))+q_{t-1}F_Y(k_1)F_Y(\widehat{k}_t^\lambda(\emptyset|N))\\ &+(1-q_{t-1})p_2^Y(I|\omega_2^B,\lambda)F_Y(\widehat{k}_t^\lambda(I|\emptyset))+(1-q_{t-1})p_2^Y(N|\omega_2^B,\lambda)F_Y(\widehat{k}_t^\lambda(N|\emptyset)).\end{aligned}$$

Together, the auxiliary mapping is defined as

$$\widehat{\mathbf{\Phi}}_t(\widehat{k}_t^\lambda,\lambda)\equiv\left(\widehat{\Phi}_t(\widehat{k}_t^\lambda,\omega_2^A,\lambda)\right)_{\{\omega_2^A\in\Omega_2^A\}},\tag{39}$$

where $\widehat{\Phi}_t(\widehat{k}_t^\lambda,\omega_2^A,\lambda)\equiv d^{-1}\big(\pi_2(\omega_2^R)\widehat{X}_t(\widehat{\mathbf{k}}_t^\lambda,\omega_2^B,\lambda)+(1-\pi_2(\omega_2^R))\widehat{Y}_t(\widehat{\mathbf{k}}_t^\lambda,\omega_2^B,\lambda)\big)$. Since $\widehat{\Phi}_t(\widehat{k}_t^\lambda,\omega_2^A,\lambda)$ is monotone in $\widehat{k}_t^\lambda$, where $\widehat{\Phi}_t(\widehat{k}_t^\lambda,\omega_2^A,\lambda)\equiv d^{-1}\big(\pi_2(\omega_2^R)\widehat{X}_t(\widehat{\mathbf{k}}_t^\lambda,\omega_2^B,\lambda)+(1-\pi_2(\omega_2^R))\widehat{Y}_t(\widehat{\mathbf{k}}_t^\lambda,\omega_2^B,\lambda)\big)$, it suffices to show that the mapping is increasing in λ, to apply the monotone comparative statics in Milgrom and Shannon (1994) to this parametrized approach.

Step 1. We first 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 2 for the auxiliary mapping in (39) given $\lambda>0$. We obtain a formula that is similar to the one from the proof for (iii) of Lemma 2 except for λ in the formula below. By taking the difference between $\widehat{k}_t^\lambda(\emptyset|I)$ and $\widehat{k}_t^\lambda(\emptyset|N)$,

$$\begin{aligned}&d(\widehat{k}_t^\lambda(\emptyset|I))-d(\widehat{k}_t^\lambda(\emptyset|N))\\ &=\pi_2(\emptyset^R)\widehat{X}_t(\widehat{\mathbf{k}}_t^\lambda,I^B,\lambda)+(1-\pi_2(\emptyset^R))\widehat{Y}_t(\widehat{\mathbf{k}}_t^\lambda,I^B,\lambda)-[\pi_2(\emptyset^R)\widehat{X}_t(\widehat{\mathbf{k}}_t^\lambda,N^B,\lambda)+(1-\pi_2(\emptyset^R))\widehat{Y}_t(\widehat{\mathbf{k}}_t^\lambda,I^B,\lambda)]\\ &=\pi_2(\emptyset^R)(1-q_{t-1})\left[F_X(\widehat{k}_t^\lambda(I|\emptyset))-F_X(\widehat{k}_t^\lambda(N|\emptyset))\right]\lambda\big(\pi_2(I^B)-\pi_2(N^B)\big)[F_Y(k_1)-F_X(k_1)]\\ &\quad+(1-\pi_2(\emptyset^R))(1-q_{t-1})\left[F_Y(\widehat{k}_t^\lambda(I|\emptyset))-F_Y(\widehat{k}_t^\lambda(N|\emptyset))\right]\lambda\big(\pi_2(I^B)-\pi_2(N^B)\big)[F_Y(k_1)-F_X(k_1)]\\ &=\frac{1}{2}\lambda(1-q_{t-1})\big(\pi_2(I^B)-\pi_2(N^B)\big)[F_Y(k_1)-F_X(k_1)]\left[\begin{array}{l}F_X(\widehat{k}_t^\lambda(I|\emptyset))-F_X(\widehat{k}_t^\lambda(N|\emptyset))\\ +F_Y(\widehat{k}_t^\lambda(I|\emptyset))-F_Y(\widehat{k}_t^\lambda(N|\emptyset))\end{array}\right],\end{aligned}$$

which establishes Lemma (iii) version for 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

$$d(\widehat{k}_t^\lambda(I|\emptyset))-d(\widehat{k}_t^\lambda(N|\emptyset))=\big(\pi_2(I^R)-\pi_2(N^R)\big)[\widehat{X}_t(\widehat{\mathbf{k}}_t^\lambda,\emptyset^B,\lambda)-\widehat{Y}_t(\widehat{\mathbf{k}}_t^\lambda,\emptyset^B,\lambda)].$$

Suppose, on the contrary, that $\widehat{k}_t^\lambda(I|\emptyset)\geq\widehat{k}_t^\lambda(N|\emptyset)$. Then, Lemma (iii) version for 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 2, which leads to (38). Then, we reach the same contradiction.

Step 2. Then for each $\omega_2^A \in \Omega_2^A$, the derivative of $\widehat{\emptyset}_t(\widehat{k}_t^{\lambda}, \omega_2^A, \lambda)$ with respect to $\lambda > 0$ yields

$$\begin{aligned}
&(1-q_{t-1})\pi_2(\omega_2^R)[-(1-\pi_2(\omega_2^B))(1-F_X(k_1)) + (1-\pi_2(\omega_2^B))(1-F_Y(k_1))]F_X(\widehat{k}_t^{\lambda}(I|\emptyset)) \\
&+(1-q_{t-1})\pi_2(\omega_2^R)[-(1-\pi_2(\omega_2^B))F_X(k_1) + (1-\pi_2(\omega_2^B))F_Y(k_1)]F_X(\widehat{k}_t^{\lambda}(N|\emptyset)) \\
&+(1-q_{t-1})(1-\pi_2(\omega_2^R))[\pi_2(\omega_2^B))(1-F_X(k_1)) - \pi_2(\omega_2^B)(1-F_Y(k_1))]F_Y(\widehat{k}_t^{\lambda}(I|\emptyset)) \\
&+(1-q_{t-1})(1-\pi_2(\omega_2^R))[\pi_2(\omega_2^B)F_X(k_1) - \pi_2(\omega_2^B)F_Y(k_1)]F_Y(\widehat{k}_t^{\lambda}(N|\emptyset)),
\end{aligned}$$

which can be rewritten as

$$\begin{aligned}
&(1-q_{t-1})\pi_2(\omega_2^R)(1-\pi_2(\omega_2^B))[F_Y(k_1) - F_X(k_1)][F_X(\widehat{k}_t^{\lambda}(N|\emptyset)) - F_X(\widehat{k}_t^{\lambda}(I|\emptyset))] \\
&-(1-q_{t-1})(1-\pi_2(\omega_2^R))\pi_2(\omega_2^B)[F_Y(k_1) - F_X(k_1)][F_Y(\widehat{k}_t^{\lambda}(N|\emptyset)) - F_Y(\widehat{k}_t^{\lambda}(I|\emptyset))].
\end{aligned} \tag{40}$$

Let's examine the above equation (40) closely for each $\omega_2^A \in \Omega_2^A$: by substituting (2),
(i) $\omega_2^A = I|\emptyset$

$$\frac{(1-q_{t-1})[F_Y(k_1) - F_X(k_1)]}{2(2 - F_X(k_1) - F_Y(k_1))} \begin{Bmatrix} (1-F_X(k_1))[F_X(\widehat{k}_t^{\lambda}(N|\emptyset)) - F_X(\widehat{k}_t^{\lambda}(I|\emptyset))] \\ -(1-F_Y(k_1))[F_Y(\widehat{k}_t^{\lambda}(N|\emptyset)) - F_Y(\widehat{k}_t^{\lambda}(I|\emptyset))] \end{Bmatrix},$$

(ii) $\omega_2^A = \emptyset|I$

$$\frac{(1-q_{t-1})[F_Y(k_1) - F_X(k_1)]}{2(2 - F_X(k_1) - F_Y(k_1))} \begin{Bmatrix} (1-F_Y(k_1))[F_X(\widehat{k}_t^{\lambda}(N|\emptyset)) - F_X(\widehat{k}_t^{\lambda}(I|\emptyset))] \\ -(1-F_X(k_1))[F_Y(\widehat{k}_t^{\lambda}(N|\emptyset)) - F_Y(\widehat{k}_t^{\lambda}(I|\emptyset))] \end{Bmatrix},$$

(iii) $\omega_2^A = \emptyset|N$

$$\frac{(1-q_{t-1})[F_Y(k_1) - F_X(k_1)]}{2(F_X(k_1) + F_Y(k_1))} \begin{Bmatrix} F_Y(k_1)[F_X(\widehat{k}_t^{\lambda}(N|\emptyset)) - F_X(\widehat{k}_t^{\lambda}(I|\emptyset))] \\ -F_X(k_1)[F_Y(\widehat{k}_t^{\lambda}(N|\emptyset)) - F_Y(\widehat{k}_t^{\lambda}(I|\emptyset))] \end{Bmatrix},$$

(iv) $\omega_2^A = N|\emptyset$

$$\frac{(1-q_{t-1})[F_Y(k_1) - F_X(k_1)]}{2(F_X(k_1) + F_Y(k_1))} \begin{Bmatrix} F_X(k_1)[F_X(\widehat{k}_t^{\lambda}(N|\emptyset)) - F_X(\widehat{k}_t^{\lambda}(I|\emptyset))] \\ -F_Y(k_1)[F_Y(\widehat{k}_t^{\lambda}(N|\emptyset)) - F_Y(\widehat{k}_t^{\lambda}(I|\emptyset))] \end{Bmatrix}.$$

Then, by $\widehat{k}_t^{\lambda}(N|\emptyset) > \widehat{k}_t^{\lambda}(I|\emptyset)$ from Step 1, if (A5) holds, we have a positive sign for all four cases.

Part 2. We show (ii) for q_{t-1} sufficiently close to 1. Since $\widehat{\Phi}_t(\widehat{k}_t^{\lambda}, \omega_2^A, \lambda)$ is continuous in λ, q_{t-1}, for each $\omega_2^A \in \Omega_2^A$, we have

$$\frac{\partial^2 \widehat{\Phi}_t(\widehat{k}_t^{\lambda}, \omega_2^A, \lambda)}{\partial \lambda \partial q_{t-1}} < 0,$$

which implies

$$q_{t-1} \to 1, \frac{\partial \widehat{\Phi}_t(\widehat{k}_t^{\lambda}, \omega_2^A, \lambda)}{\partial \lambda} \to 0 \text{ for each } \lambda > 0 \text{ and } q_{t-1} > \frac{1}{2}.$$

This establishes that a homogeneous mapping in (8) and a heterogeneous mapping in (12) are uniformly close to each other for a sufficiently high q_{t-1}. ■

Proof of Lemma 4. *(i)* First, by combining Lemma 3 and Proposition 3, with (16) and (20), a location equilibrium as given by Definition 1 can be simplified such that ℓ_t is a location equilibrium if (i) for each $\omega_2 \in \Omega_2$,

$$\begin{aligned}
\ell_t^B(\omega_2) &= \begin{cases} E & \text{if } \Delta P_t(\omega_2, \ell_t)[\mathbf{1}_{\{\omega_2 \in \Omega_2^{B+}\}} - \mathbf{1}_{\{\omega_2 \in \Omega_2^{B-}\}}] > 0, \\ W & \text{if } \Delta P_t(\omega_2, \ell_t)[\mathbf{1}_{\{\omega_2 \in \Omega_2^{B+}\}} - \mathbf{1}_{\{\omega_2 \in \Omega_2^{B-}\}}] < 0, \end{cases} \\
\ell_t^R(\omega_2) &= \begin{cases} E & \text{if } \Delta P_t(\omega_2, \ell_t)[\mathbf{1}_{\{\omega_2 \in \Omega_2^{R+}\}} - \mathbf{1}_{\{\omega_2 \in \Omega_2^{R-}\}}] < 0, \\ W & \text{if } \Delta P_t(\omega_2, \ell_t)[\mathbf{1}_{\{\omega_2 \in \Omega_2^{R+}\}} - \mathbf{1}_{\{\omega_2 \in \Omega_2^{R-}\}}] > 0, \end{cases}
\end{aligned} \tag{41}$$

and (ii) $P_t^B(\omega_2, \ell_t^B) = \mathbb{E}\big[\mathbf{1}_{\{\ell_t^B(\widetilde{\omega}_2)=E\}} \mid \omega_2\big]$ and $P_t^R(\omega_2, \ell_t^R) = \mathbb{E}\big[\mathbf{1}_{\{\ell_t^R(\widetilde{\omega}_2)=E\}} \mid \omega_2\big]$. Since $\mathbf{1}_{\{\omega_2 \in \Omega_2^{B+}\}} - \mathbf{1}_{\{\omega_2 \in \Omega_2^{B-}\}} = -[\mathbf{1}_{\{\omega_2 \in \Omega_2^{R+}\}} - \mathbf{1}_{\{\omega_2 \in \Omega_2^{R-}\}}]$, for all ω_2, we have $\ell_t^B(\omega_2) = \ell_t^R(\omega_2)$.

(ii) Given $\ell_t^B(\omega_2) = \ell_t^R(\omega_2)$,

$$\begin{aligned}
\Delta P_t(\omega_2, \ell_t) &= \mathbb{E}\big[\mathbf{1}_{\{\ell_t^B(\widetilde{\omega}_2)=E\}} \mid \omega_2\big] - \mathbb{E}\big[\mathbf{1}_{\{\ell_t^R(\widetilde{\omega}_2)=E\}} \mid \omega_2\big] \\
&= (2q_{t-1}-1)p(I|\omega_2^B)\mathbf{1}_{\{\ell_t^g(I,\emptyset)=E\}} - (2q_{t-1}-1)p(N|\omega_2^R)\mathbf{1}_{\{\ell_t^g(\emptyset,N)=E\}} \\
&\quad + (2q_{t-1}-1)p(N|\omega_2^B)\mathbf{1}_{\{\ell_t^g(N,\emptyset)=E\}} - (2q_{t-1}-1)p(I|\omega_2^R)\mathbf{1}_{\{\ell_t^g(\emptyset,I)=E\}},
\end{aligned}$$

which results in (21) given $p(I|\omega_2^R) = 1 - p(N|\omega_2^R)$ and $p(I|\omega_2^B) = 1 - p(N|\omega_2^B)$. We divide the proof into two cases.

Case 1. Consider Ω_2^{B+} and note that for each $\omega_2 = (\omega_2^B, \omega_2^R) \in \Omega_2^{B+}$, $p(N|\omega_2^B) < p(N|\omega_2^R)$ from (10). Suppose $\mathbf{1}_{\{\ell_t(I,\emptyset)=E\}} \neq \mathbf{1}_{\{\ell_t(\emptyset,N)=E\}}$. First, suppose $\mathbf{1}_{\{\ell_t(I,\emptyset)=E\}} = 1$, $\mathbf{1}_{\{\ell_t(\emptyset,N)=E\}} = 0$. Since $(\emptyset, N) \in \Omega_2^{B+}$ but $\mathbf{1}_{\{\ell_t(\emptyset,N)=E\}} = 0$ (choosing W), from (21), for $\omega_2 = (\emptyset, N)$, 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. Now, suppose $\mathbf{1}_{\{\ell_t(I,\emptyset)=E\}} = 0$, $\mathbf{1}_{\{\ell_t(\emptyset,N)=E\}} = 1$. Since $(\emptyset, N) \in \Omega_2^{B+}$ but $\mathbf{1}_{\{\ell_t(\emptyset,N)=E\}} = 1$ (choosing E), from (21), for $\omega_2 = (\emptyset, N)$, 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.

Case 2. Consider Ω_2^{B-} and note that for each $\omega_2 = (\omega_2^B, \omega_2^R) \in \Omega_2^{B+}$, $p(N|\omega_2^B) > p(N|\omega_2^R)$ from (10). Suppose $\mathbf{1}_{\{\ell_t(N,\emptyset)=E\}} \neq \mathbf{1}_{\{\ell_t(\emptyset,I)=E\}}$. First, suppose $\mathbf{1}_{\{\ell_t(N,\emptyset)=E\}} = 1$, $\mathbf{1}_{\{\ell_t(\emptyset,I)=E\}} = 0$. Since $(N, \emptyset) \in \Omega_2^{B-}$ but $\mathbf{1}_{\{\ell_t(N,\emptyset)=E\}} = 1$ (choosing E), from (21), for $\omega_2 = (N, \emptyset)$, 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. Now, suppose $\mathbf{1}_{\{\ell_t(N,\emptyset)=E\}} = 0$, $\mathbf{1}_{\{\ell_t(\emptyset,I)=E\}} = 1$. Since $(N,\emptyset) \in \Omega_2^{B-}$ but $\mathbf{1}_{\{\ell_t(N,\emptyset)=E\}} = 0$ (choosing W), from (21), for $\omega_2 = (N,\emptyset)$, 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.

(iii) Given the result (ii), if for each $\omega_2 \in \Omega_2^{g+}$, $\omega_2' \in \Omega_2^{g-}$, $\ell_t^g(\omega) = \ell_t^g(\omega')$, we have $\Delta P_t(\omega_2,\ell_t) = 0$, which contradicts what we suppose $\Delta P_t(\omega_2,\ell_t) \neq 0$. ■

Proof of Proposition 4. First, note that $(2q_{t-1}-1) > 0$ from Lemma 1. Given the formula in (23), if $1 - N_2(\omega_2) > 0$ for all $\omega_2 \in \Omega_2$, there exists a binary splitting location equilibrium. Now, suppose a binary splitting location equilibrium. From FOSD in (A2), $p_2(N|\omega_2^B)$ is strictly decreasing in $\pi_2(\omega_2^g)$, so for $(\omega_2^B,\omega_2^R) = (N,\emptyset)$ or $(\omega_2^B,\omega_2^R) = (\emptyset,N)$ and for $(\widetilde{\omega}_2^B,\widetilde{\omega}_2^R) = (I,\emptyset)$ or $(\widetilde{\omega}_2^B,\widetilde{\omega}_2^R) = (\emptyset,I)$,

$$N_2(\omega_2^B,\omega_2^R) = p_2(N|\omega_2^B) + p_2(N|\omega_2^R) > p_2(N|\widetilde{\omega}_2^B) + p_2(N|\widetilde{\omega}_2^R) = N_2(\widetilde{\omega}_2^B,\widetilde{\omega}_2^R),$$

where recall

$$\begin{aligned}
&p_2(N|\omega_2^B) + p_2(N|\omega_2^R) \\
&= \pi_2(\omega_2^B)F_X(k_1) + (1-\pi_2(\omega_2^B))F_Y(k_1) + \pi_2(\omega_2^R)F_X(k_1) + (1-\pi_2(\omega_2^R))F_Y(k_1).
\end{aligned}$$

This implies that $1 - N_2(\omega_2) > 0$ for all $\omega_2 \in \Omega_2$ if and only if $1 - N_2(N,\emptyset) > 0$ in Proposition 4. Now, suppose that $1 - N_2(\omega_2) \leq 0$ for some $\omega_2 \in \Omega_2$. If $1 - N_2(\omega_2) < 0$, we have a contradiction with Lemma 4 and (23), and if $1 - N_2(\omega_2) = 0$, there can other location equilibria due to the indifference. ■

Proof of Proposition 5. *Part 1.* First, using γ, the addition of systematic polarization changes the formula in (23) to

$$\Delta P_t(\omega_2,\ell_t) = \begin{cases} \alpha(2q_{t-1}-1)[1-N_2(\omega_2)] + (1-\alpha)\gamma\Delta\mathcal{P}_{t-1} & \text{if } \Delta P_t(\omega_2,\ell_t) > 0, \\ -\alpha(2q_{t-1}-1)[1-N_2(\omega_2)] + (1-\alpha)\gamma\Delta\mathcal{P}_{t-1} & \text{if } \Delta P_t(\omega_2,\ell_t) < 0. \end{cases} \tag{42}$$

Note that q_{t-1} from (4) can be rewritten as:

$$\begin{aligned} q_{t-1} &\equiv \frac{\mathcal{P}^B_{t-1}+\mathcal{P}^R_{t-1}-2\mathcal{P}^B_{t-1}\mathcal{P}^R_{t-1}}{\left(\mathcal{P}^B_{t-1}+\mathcal{P}^R_{t-1}\right)\left(2-\mathcal{P}^B_{t-1}-\mathcal{P}^R_{t-1}\right)}. \\ &= \frac{1}{2-(\mathcal{P}^B_{t-1}+\mathcal{P}^R_{t-1})} - \frac{(\mathcal{P}^B_{t-1}+\mathcal{P}^R_{t-1})^2-\Delta\mathcal{P}^2_{t-1}}{2(\mathcal{P}^B_{t-1}+\mathcal{P}^R_{t-1})(2-\mathcal{P}^B_{t-1}-\mathcal{P}^R_{t-1})} \\ &= \frac{1}{2}+\frac{\Delta\mathcal{P}^2_{t-1}}{2A_{t-1}(2-A_{t-1})}. \end{aligned}$$

Then, by substituting the above into (42), we obtain (25).

Part 2. Consider the belief dynamics (25).

(i) For $\Delta P_t(\omega_2,\ell_t)<0$, the square function yields a positive horizontal intercept $\frac{(1-\alpha)\gamma}{\alpha}\frac{A_{t-1}(2-A_{t-1})}{1-N_2(\omega_2)}$ to make (25) zero (see the right panel of Figure 5). If $\Delta\mathcal{P}_{t-1}<\frac{(1-\alpha)\gamma}{\alpha}\frac{A_{t-1}(2-A_{t-1})}{1-N_2((N,\emptyset))}$ where we choose $\omega_2=(N,\emptyset)$ for $N_2(\omega_2)$, an equilibrium with $\Delta P_t(\omega_2,\ell_t)<0$ cannot arise since the square function yields a positive value for $\omega_2=(N,\emptyset)$, contradicting $\Delta P_t(\omega_2,\ell_t)<0$. On the other hand, if $\Delta\mathcal{P}_{t-1}>\frac{(1-\alpha)\gamma}{\alpha}\frac{A_{t-1}(2-A_{t-1})}{1-N_2((N,\emptyset))}$. both $\Delta P_t(\omega_2,\ell_t)>0$ and $\Delta P_t(\omega_2,\ell_t)<0$ equilibria can arise. That is, the possible sign of the period t equilibrium belief depends on whether the period $t-1$ actual population difference $\Delta\mathcal{P}_{t-1}$ is large or small. The only robust equilibrium belief is that $\Delta\mathcal{P}_{t-1}>0\,(<0)$ yields $\Delta P_t(\omega_2,\ell_t)>0\,(<0)$, which is corresponding to a Markov location strategy (26). With a Markov location strategy, the belief dynamics (25) changes into

$$\Delta P_t(\omega_2,\ell_t)=\begin{cases}\alpha\frac{1-N_2(\omega)}{A_{t-1}(2-A_{t-1})}\Delta\mathcal{P}^2_{t-1}+(1-\alpha)\gamma\Delta\mathcal{P}_{t-1} & \text{if } \Delta\mathcal{P}_{t-1}>0,\\ -\alpha\frac{1-N_2(\omega)}{A_{t-1}(2-A_{t-1})}\Delta\mathcal{P}^2_{t-1}+(1-\alpha)\gamma\Delta\mathcal{P}_{t-1} & \text{if } \Delta\mathcal{P}_{t-1}<0,\end{cases} \tag{43}$$

where note that $\Delta P_t(\omega_2,\ell_t)$ only depends on the sign of $\Delta\mathcal{P}_{t-1}$, unlike (25).

(ii) It is immediate that $\Delta\mathcal{P}_{t-1}=0$ implies $\Delta P_t(\omega_2,\ell_t)=0$. Consider $\Delta P_t(\omega_2,\ell_t)=0$ and suppose $\Delta\mathcal{P}_{t-1}\neq 0$. Then, we have a contradiction given (43). ■

Proof of Lemma 5. From an outside theorist's point of view, $p(N|\omega_2^B)=F_B(k_1)$ and $p(N|\omega_2^R)=F_R(k_1)$, and the summation of them yields (30). Then, one can show that A_t converges to $1-F_B(k_1)+F_R(k_1)$ such that $A_t>A_{t-1}$ for all $t=1,2,....$ if $A_0<1-F_B(k_1)+F_R(k_1)$, whereas $A_t<A_{t-1}$ for all $t=1,2,....$ if $A_0>1-F_B(k_1)+F_R(k_1)$ since from (30), we have

$$A_t-A_{t-1}=\begin{cases}\alpha[1-F_B(k_1)+F_R(k_1)-A_{t-1}]>0 & \text{if } A_0<1-F_B(k_1)+F_R(k_1),\\ \alpha[1-F_B(k_1)+F_R(k_1)-A_{t-1}]<0 & \text{if } A_0>1-F_B(k_1)+F_R(k_1).\end{cases}$$

For any combination of (F_B,F_R), if $A_0<1-|F_B(k_1)-F_R(k_1)|$, A_{t-1} strictly increases, whereas $A_0>1+|F_B(k_1)-F_R(k_1)|$, A_{t-1} strictly decreases. Both cases make $A_{t-1}(2-A_{t-1})$ increase, so x^*_{t-1} through (29). ■

Proof of Proposition 6. The real dynamics in (27) can be rewritten as:

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

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 A_0 satisfying Lemma 5, $x_t^*>x_{t-1}^*$ 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

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

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

Proof of Proposition 7. First, we show that if $\alpha\beta_t\Delta\mathcal{P}_t+(1-\alpha)\gamma>1$ then $\alpha\beta_{t-1}\Delta\mathcal{P}_{t-1}+(1-\alpha)\gamma>1$ for all $t=1,2,.....$ Suppose, on the other hand, that there exists t such that $\alpha\beta_{t-1}\Delta\mathcal{P}_{t-1}+(1-\alpha)\gamma\leq 1$ and $\alpha\beta_t\Delta\mathcal{P}_t+(1-\alpha)\gamma>1$. Then, $\Delta\mathcal{P}_t\leq\Delta\mathcal{P}_{t-1}$. And $\beta_t<\beta_{t-1}$ by the proof of Lemma 5. Hence,

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

which is a contradiction. We define f_{t-1} in (27) in the positive domain and that in the negative domain, respectively, as

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

and their inverse functions as $h_{t-1}^{(+)}(x)$ and $h_{t-1}^{(-)}(x)$.

Part (i) $1-\mathcal{N}_2>0$. Find $\widehat{t}$ such that $\widehat{t}\equiv\max\{t:x_t^{(+)}\leq 1\}$. Then, given $\widehat{t}$, we obtain $\Delta\mathcal{P}_{\widehat{t}}=h_{\widehat{t}}^{(+)}(1)$, which yields

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

Finally, denote by $\widehat{x}$ a critical level satisfying a relatively high difference in Proposition 3, and then $x^{\dagger}=\max\{x',\widehat{x}\}$.

Part (ii) Consider a sequence of functions such that $f_0^{(+)},f_1^{(-)},....$ with their corresponding fixed points such as $x_0^{(+)},x_1^{(-)},.....$ Now denote

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

Find $\widehat{t}$ such that $\widehat{t}\equiv\max\Big\{t:\big|x_t^a\big|\leq 1\Big\}$. 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, we find

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

As in Part (i), $x^{\dagger\dagger}=\max\{x'',\widehat{x}\}$. 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 5, $A_{t-1} > 1$, we have $\alpha + (1-\alpha)A_{t-1} > 1$. Then, given $1 - F_B(k_1) + F_R(k_1) > 1$, by comparing $F_B < F_R$ with $F_B = F_R$,

$$\alpha[1 - F_B(k_1) + F_R(k_1)] + (1-\alpha)A'_{t-1} > \alpha + (1-\alpha)A_{t-1},$$

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 $\mathcal{N}'_2 = F_X(k_1) + F_Y(k_1)$ for the different distributions and $\mathcal{N}_2 = F_Y(k_1) + F_Y(k_1)$ for the same distribution. By FOSD between F_X and F_Y, we have $1 - \mathcal{N}'_2 > 1 - \mathcal{N}_2$. Overall, consider (28), 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 5, $A_{t-1} > 1$, we have $\alpha + (1-\alpha)\big(\mathcal{P}^B_{t-1} + \mathcal{P}^R_{t-1}\big) > 1$. Then, given $1 - F_B(k_1) + F_R(k_1) < 1$, by comparing $F_B > F_R$ with $F_B = F_R$, we have

$$\alpha[1 - F_B(k_1) + F_R(k_1)] + (1-\alpha)A''_{t-1} < \alpha + (1-\alpha)A_{t-1},$$

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 $\mathcal{N}''_2 = F_X(k_1) + F_Y(k_1)$ for the different distributions and $\mathcal{N}_2 = F_X(k_1) + F_X(k_1)$ for the same distribution. By FOSD between F_X and F_Y, we have $1 - \mathcal{N}''_2 < 1 - \mathcal{N}_2$. Overall, consider (28), 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 (27). We denote f_{t-1} given $F_B = F_R$ and say that a function g_{t-1} uniformly dominates f_{t-1} if $g_{t-1}(x) > f_{t-1}(x)$ for all x. 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

Baccara, M. and Yariv, L. (2016), Choosing peers: Homophily and polarization in groups. *Journal of Economic Theory* 165, 152–178.

Becker, G.S. (1973), A theory of marriage: Part i, *Journal of Political Economy* 81, 89–125.

Board, S. (2009), Monopolistic group design with peer effects, *Theoretical Economics* 4, 813–846.

Burdett, K. and Coles, M.G., Marriage and class, *Quarterly Journal of Economics*, 112, 141–168.

Calvó-Armengol, A., De Martí, J. and Prat, A. (2015), Communication and influence. *Theoretical Economics* 10, 649–690.

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

Currarini, S., Jackson, M.O and Pin, P., 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.

Galeotti, A., Ghiglino, C. and Squintani, F. (2013), Strategic information transmission networks, *Journal of Economic Theory* 148, 1751–1769.

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

Hoppe, H. C., Moldovanu, B. and Ozdenoren, E. (2011), Coarse matching with incomplete information, *Economic Theory* 47, 75–104.

Huckfeldt, R.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.

Kandori, M. (2002) Introduction to repeated games with private monitoring, *Journal of Economic Theory* 102, 1–15.

Kets, W. and Sandroni, A. (2019), A Belief-based theory of homophily, *Games and Economic Behavior* 115, 410–435.

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

Legros, P. and Newman, A. F. (2002), Monotone matching in perfect and imperfect worlds, *Review of Economic Studies* 69, 925–942.

Legros, P. and Newman, A. F. (2007), Beauty is a beast, frog is a prince: Assortative matching with nontransferabilities, *Econometrica* 75, 1073–1102.

Levy, G. and Razin, R. (2019), Echo chambers and their effects on economic and political outcomes, *Annual Review of Economics* 11, 303–328.

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

Morgan, P. (1995), A model of search, coordination, and market segmentation, Manuscript, SUNY-Buffalo.

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.

Peski, M. (2008), Complementarities, group formation and preferences for similarity, Working paper, University of Toronto.

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.

Shimer, R. and Smith, L. (2000), Assortative matching and search, *Econometrica* 68, 343–369.

Smith, L. (2006), The marriage model with search frictions, *Journal of Political Economy* 114, 1124–1144.

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

Tajfel, Henri (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, 415–38.