Preference Evolution under Partner Choice*

Ziwei Wang† Jabin Wu‡

March 15, 2025

Abstract

We present a model that investigates preference evolution with endogenous matching. In the short-run, individuals' subjective preferences influence partner selection and behavior in social interactions, which affects their material payoffs. These payoffs, in turn, affect how preferences evolve in the long-run. To properly model the "match-to-interact" process, we combine stable matching and equilibrium concepts. Our analysis shows that endogenous matching gives rise to the evolutionary stability of a class of preferences that exhibit both homophily and efficiency. Such preferences stand out in the evolutionary process because they are able to force positive assortative matching and efficient play. Under incomplete information, a strong form of homophily, termed *parochialism*, is necessary for a preference to prevail in evolution, because stronger incentives are required to engage in self-sorting with information friction.

Key Words: Preference evolution, stable matching, evolutionary stability, matching with incomplete information, homophily, parochialism.

1 Introduction

In contemporary economic analysis, decision makers' preferences are commonly taken as exogenously given and fixed. However, preferences themselves can be the products of a

*We are grateful to the co-editor and four anonymous referees for their valuable suggestions that significantly improved the paper. We thank Ingela Alger, Yi-Chun Chen, Jeffrey Ely, Wei He, Gaoji Hu, Luis Izquierdo, Segismundo Izquierdo, Sam Jindani, Jinwoo Kim, Ce Liu, Qingmin Liu, Jonathan Newton, Fanqi Shi, Xiang Sun, Yifei Sun, Ina Taneva, Qianfeng Tang, Yi Tong, Matthijs van Veelen, Jögen Weibull, Xi Weng, Yiqing Xing, Nate Yoder, Hanzhe Zhang, Kun Zhang, Jidong Zhou and participants at various conferences and seminars for constructive comments. Wang acknowledges financial support from the National Natural Science Foundation of China (No. 72403188). The usual disclaimer applies.

†Peking University and Wuhan University, zwang.econ@gmail.com

‡University of Oregon, jwu5@uoregon.edu

lengthy evolutionary process. An important question arises regarding why certain preferences persist while others dissipate throughout human history. Güth and Yaari (1992) and Güth (1995) introduce the "indirect evolutionary approach," a useful theoretical framework for understanding preference evolution: preferences dictate behavior, behavior determines fitness success, and fitness success, in turn, regulates how preferences evolve. A preference type is considered evolutionarily stable if, when predominant in a population, it can resist invasions from alternative preference types. See recent surveys by Alger and Weibull (2019) and Alger (2022).[1] In most works that align with this approach, behavior refers to the choices made in some two-player game by a population of individuals who are paired according to some exogenous random matching process.[2] Nevertheless, they neglect to consider a crucial aspect of behavior: preferences not only shape individuals' choices in the underlying game after they are matched but also determine their matching patterns in the first place, i.e., how people choose their partners and get paired with one another. Without acknowledging the role of preferences in determining matching, and how matching in turn shapes preferences, the indirect evolutionary approach remains incomplete.

The objective of this paper is to propose a model of preference evolution, which formally incorporates endogenous matching given an arbitrary space of preference types, any underlying two-player game with finite strategies, and any information available to the individuals in the population. Our approach to model the match-to-play-game process follows the works of Jackson and Watts (2010) and Garrido-Lucero and Laraki (2021) by integrating the non-cooperative concept of equilibrium play and the cooperative notion of stable matching. On one hand, equilibrium ensures that the play between two matched individuals is self-enforcing once the matching is formed. On the other hand, stable matching captures the idea that unsatisfied pairs of individuals can communicate and jointly deviate to form new pairs in a credible manner. Alternatively, endogenous matching can be modeled in a completely non-cooperative fashion; Nevertheless, this approach may result in a complex and intractable extensive form that is difficult to analyze, and the solution can be sensitive to the assumed protocol. Thus, we adopt the protocol-free cooperative approach, which is robust to the details of the extensive form.[3]

We begin the paper with a complete information benchmark where preference types are observable. To investigate the stability of matching outcomes in this context, we adapt the

[1] See also Robson and Samuelson (2011) for a critical assessment of the approach.

[2] A few papers consider multilateral games (Lehmann et al., 2015; Alger and Weibull, 2016; Alger et al., 2020) or all individuals play "the field" (Lahkar, 2019; Bandhu and Lahkar, 2023). In the former case, matching is also assumed to be exogenous, while in the latter case, matching plays no role.

[3] While stable matching can be viewed as a reduced-form limit of some frictionless dynamic processes of partnership formation, the convergence of such dynamics is not guaranteed. More importantly, in reality, frictions naturally arise during these processes. Therefore, using the concept of stable matching serves as an initial step in modeling endogenous partner choice in the evolution of preferences.

concept of *Nash stability* by Garrido-Lucero and Laraki (2021) to our setting. Specifically, Nash stability requires that every matched pair of individuals plays a Nash equilibrium, termed internal stability, and that no unmatched pair can coordinate on a Nash equilibrium that benefits them both, termed external stability.

We identify a class of preference types that exhibit distinctive characteristics. First, these preference types display a form of plasticity[4] that we refer to as *homophily*.[5] We distinguish two different forms of homophily: weak homophily, where an individual derives additional utility from interacting with another individual of the same type, and strong homophily, where an individual exclusively derives utility from interacting with another individual of the same type. We refer to the latter preference type as *parochial* because it reflects the state of mind whereby such individuals narrowly focus on interactions among themselves rather than considering the wider population that includes different types of agents.[6] Note that such individuals never remain matched with opponents of different types in a stable outcome (assuming types are observable).

Second, these preference types must induce efficient play that maximizes the material payoff of both players in the underlying game. There are two different ways to interpret the preference for efficiency. On the one hand, it can be viewed as a "disinterested" preference, where an individual is solely focused on the social objective of maximizing the total material payoff. This perspective is in line with Utilitarianism, as the strategy chosen by an individual with a preference for efficiency, conditional on her opponent's behavior, is considered morally right since it maximizes the welfare of the pair (Mill, 1863). On the other hand, the preference for efficiency can also be seen as a form of altruism, where an individual places equal importance on her own material payoff and that of her opponent. Starting from the seminal works of Hamilton (1964a,b), it is generally understood in the theoretical biology literature that the maintenance of altruism in evolution, whether the solutions are based on kinship, reciprocity, or group selection, depends on assortative matching.

The rationales for these preferences to be evolutionarily stable are as follows: First, homophily or parochialism fosters positive assortative matching, ensuring that all agents carrying this preference type are matched with one another. Second, a preference for efficiency shifts these agents' incentives in the material game, making them play an efficient strategy profile (e.g. cooperation in a prisoner's dilemma) as a Nash equilibrium. Finally, playing the

[4]Plasticity refers to situations where an individual has preference over not only the strategy profiles in the underlying game but also the opponent's preference type. Works in preference evolution that consider plasticity include Sethi and Somanathan (2001), Herold and Kuzmics (2009) and Alger and Lehmann (2023).

[5]In the network literature, it is common to assume that players have the homophilous preference (a direct preference for associating with similar others). see Jackson (2014) for a survey.

[6]In the literature, various interpretations have been given to parochialism. For example, Bernhard et al. (2006) define it as a preference for favoring one's own group members. Choi and Bowles (2007) define it as hostility toward other groups.

efficient strategy profile among themselves guarantees a higher average fitness than any other types that do not play efficiently.

Next, we turn our attention to the case of incomplete information in which individuals' preference types are their private information. Following the recent literature on stable matching with incomplete information (Liu et al., 2014; Liu, 2020; Chen and Hu, 2023; Wang, 2022), we develop a stability concept called Bayes-Nash stability in our model. It requires that each matched pair plays a Bayes-Nash equilibrium (internal stability), and that there is no incomplete information blocking pair (external stability). Intuitively, an incomplete information blocking pair exists when (1) two individuals agree on a rematching proposal that specifies how each side is supposed to play in the deviation, and (2) both sides strictly benefit from the deviation given that the proposal will be honored. Notice that such blocking may be accompanied by information revelation, so the requirement of external stability imposes restrictions on the form of incomplete information in a stable outcome. This highlights that information is an *endogenous* variable in our setting. We show that individuals with the stronger *parochial efficient* preference type are able to cooperate with each other through rematching and force efficient play even in the presence of incomplete information.[7] Consequently, the parochial efficient preference type can ensure the highest average fitness and thus resist invasion of any other type that does not always induce efficient play. We also show through an example that the weaker homophilic efficient type may not resist invasion (Example 5), and formally prove that it is indeed evolutionarily *unstable* for a large class of material games. The intuition is as follows: Although the homophilic efficient types prefer to play with their own type, their utility still depends on the behavior of the opponent when matched with another type; therefore, they may be reluctant to block an unfavorable matching outcome if they cannot distinguish an (unobservable) alien type that tends to minimize total material payoffs.

Fundamentally, the mechanisms behind our model's predictions are two-fold: type-identification and commitment. Type-identification is achieved through the matching protocol combined with parochialism. The opportunities (and incentives) to block the current matching outcome enable the parochial efficient preference type to be partially revealed in a stable matching outcome due to its specific nature. The commitment power is internalized for the parochial efficient preference type (parochialism can be weakened to homophily under complete information), with which agents of this type are dedicated to matching and cooperating only with others of the same type. Frank (1987) was the first to highlight that commitment

[7]While our focus is on the stable outcome rather than explicitly modeling the dynamic matching process converging to it, we can imagine that individuals with the parochial efficient preference type engage in a self-sorting process that involves information unraveling. We demonstrate that whenever individuals with unobservable types are not matched among themselves playing an efficient strategy pair, a profitable rematching opportunity arises. Carrying out such deviations can reveal more information to the population.

and the ability to identify committing agents serve as important driving forces of evolution. Recent works by Akdeniz and van Veelen (2021, 2023), along with our model, can be seen as a renaissance of Frank's insight.

In traditional economic models, individuals are generally assumed to be self-interested and strive to maximize their own material payoffs. Therefore, it is important to investigate whether selfishness can be evolutionarily stable in the context of preference evolution. To this end, we examine two types of selfish preferences that exhibit homophily: the weaker *homophilic selfish* and the stronger *parochial selfish* types. Our findings indicate that with complete information, for these types to be stable, additional conditions must be met for the material game. In particular, an important subset of Nash equilibria, which we call the set of *loser-best Nash equilibria*, must be efficient.

As previous works in the literature suggest, selfishness may be favored by natural selection under incomplete information when the matching process is assumed to be exogenous (Ely and Yilankaya, 2001; Ok and Vega-Redondo, 2001; Dekel et al., 2007). In contrast, our results imply that with endogenous matching, a preference for efficiency still dominates selfishness, even when incomplete information is present. In fact, incomplete information makes it even more challenging for selfish types to be evolutionarily stable. Specifically, we show that *all* Nash equilibria in the material game must be efficient to ensure the evolutionary stability of the parochial selfish type, which is a more stringent condition than with complete information.

1.1 Related Literature

In what follows, we review the literature related to the current work. The idea of the indirect evolutionary approach has already been proposed by several earlier papers including Becker (1976), Hirshleifer (1977), Rubin and Paul (1979), and Frank (1987) before it is formally named.[8] Frank (1987) is perhaps the first paper that considers endogenous matching under incomplete information. His model imposes an exogenous information structure that induces positive assortative matching and the agents interact only once after they are matched. On the contrary, we adopt the notion of stable matching as a reduced-form limit of a repeated match-to interact process, which endogenizes the information structure.

The concern of how incomplete information affects preference evolution in a random matching environment dates back to Güth and Kliemt (1998), who demonstrates that conditioner cooperators cannot survive when preference types are unobservable because

[8] Additional subsequent works include Robson (1990), Ockenfels (1993), Ellingsen (1997), Bester and Güth (1998), Güth and Kliemt (1998), Fershtman and Weiss (1998), Huck and Oechssler (1999) McNamara et al. (1999), Bolle (2000), Koçkesen et al. (2000), Possajennikov (2000), Sethi and Somanathan (2001), Van Veelen (2006), Heifetz et al. (2007a,b), Akçay et al. (2009), Alger (2010), Alger and Weibull (2010, 2012), Carvalho et al. (2023), and Avataneo et al. (2025), among others.

they cannot behave differently according to their opponents' types. Dekel et al. (2007) formalize the idea in a fairly general setting.They find that when preference is perfectly observable, efficiency is the driving force behind the selection of behavior. When preference types are completely unobservable, selfishness is instead evolutionarily stable. They also consider the intermediate case that the individuals' types are partially observable, where the degree of observability is exogenously given. They find that efficiency force matters for any positive degree of observability; only when preferences are completely unobservable does this force disappear. Herold and Kuzmics (2009) extends Dekel et al. (2007) to allow for plasticity in a random matching environment. They show that when plasticity is incorporated, discriminating types are evolutionarily stable under (almost) complete information. In our model, we do not impose any exogenous information structure. Instead, observability is endogenized by the matching process. We show that efficiency force joint with parochialism, as a form of discrimination, prevails regardless of the initial degree of observability before matching because it induces information revelation and assortative matching.

Alger and Weibull (2013) consider a preference evolution model with incomplete information and exogenous assortative matching. That is, individuals with the same preference types are matched with higher probability than those with different preference types. They establish that, contrary to previous findings, a preference type called *homo-moralis*, which concerns both materialistic goals and moral values, is evolutionarily stable. In the most extreme case where there is positive assortative matching, the *Kantian* preference type, which aligns with the philosophy of Kant (1785), becomes evolutionarily stable. It is worth noting that a *Kantian* individual's dominant strategy corresponds to the symmetric efficient strategy profile in the underlying game. Newton (2017) extends Alger and Weibull (2013) by subjecting matching's degree of assortativity to evolutionary pressure. He demonstrates that the *Kantian* preference coupled with homophily defined on the matching level can survive. In our model, homophily is rather defined on the more primitive preference level and we further endogenize assortativity to an individual level by employing the concept of stable matching. In addition, Wu (2019) correlates observability of preference types with assortativity of matching exogenously. In contrast, our paper endogenizes such correlation.

Different ways of modeling match-to-play-game are proposed in the literature. Ely (2002) and Mailath et al. (1997) consider models where the interaction structure is endogenized by locational choices. Starting from Jackson and Watts (2002), a growing literature endogenizes interaction structure via network formation (Goyal and Vega-Redondo, 2005; Hojman and Szeidil, 2006; Staudigl and Weidenholzer, 2014; Bilancini and Boncinelli, 2018; Cui and Shi, 2021; Cui and Weidenholzer, 2021). Dynamic partner choice models have been considered by Frank (1987), Wilson and Dugatkin (1997), McNamara et al. (2008), Fujiwara-Greve and Okuno-Fujiwara (2009), Izquierdo et al. (2010, 2014, 2021), and Graser et al. (2024)

The general takeaway from the above-described strands of literature is that when people have enough freedom to choose both whom they interact with and action in the underlying games, efficiency arises. The concepts of stable matching we develop in this paper implicitly assume that the matching process is frictionless. In addition, Gintis et al. (2001) and Hopkins (2014) among others, use the costly signaling theory to model endogenous matching under incomplete information. Nax and Rigos (2016) and Wu (2017) consider models in which matching's degree of assortativity is determined through political processes.

2 Population, Strategies and Preference Types

Consider a continuum of agents constituting a population who are matched in pairs to engage in symmetric two-person simultaneous game Γ with a common strategy set X. We assume X is finite and allow the agents to choose from the set of mixed strategies denoted by $\mathcal{X} = \Delta(X)$. An agent playing pure strategy $x \in X$ against another agent playing pure strategy $y \in X$ receives a **material payoff** (or **fitness**) $\pi(x, y)$, where $\pi : X^2 \to \mathbb{R}$. The payoff function π is naturally extended to the domain of mixed strategy profiles $\mathcal{X}^2$. Because all agents have to be matched and play the material game in our model (i.e. there is no outside option), we normalize the material payoff function so that $\pi(x, y) \geqslant 0$ for all $(x, y) \in X^2$ without loss of generality.[9]

Write Θ for the set of **preference types** an agent can possess. Each preference type $\theta \in \Theta$ defines a utility function (and its affine transformations) $u_\theta : X^2 \times \Theta \to \mathbb{R}$, which depends on the pure strategies played by the pair and the matched partner's preference type. For example, $u_\theta(x, y, t)$ denotes the utility of an agent with preference type θ playing pure strategy x against another agent with preference type t playing pure strategy y. For each $\theta \in \Theta$, u_θ is naturally extended to the domain $\mathcal{X}^2 \times \Theta$. Assume Θ is rich enough so that any utility function is possessed by some preference type.[10] Our specification of the utility function is more general than those typically considered in the literature on preference evolution, as we allow it to depend on the preference type of the matched partner. This dependency potentially makes an individual less exploitable by others with different preference types, a force that becomes even more crucial when partner choice is endogenous. When u_θ is non-constant on Θ, we say type θ has **plastic** preferences.

We make two remarks. First, we focus on material games with a finite common strategy set X, as in Dekel et al. (2007). Although our analysis readily extends to games with a general

[9] Alternatively, we can consider an environment where individuals have the option to stay unmatched, but if so, they receive a material payoff lower than the minimum payoff they could obtain by interacting with a partner. In this case, all of our analysis and results still hold.

[10] For example, we can let Θ be the canonical type space constructed in Gul and Pesendorfer (2016); see also Herold and Kuzmics (2009).

topological strategy space once we impose suitable assumptions, the central insights of the paper remain unchanged. Second, we impose no relation between u_θ and the material payoff function π. Special examples include the **selfish** type that only cares about the material payoff, i.e. $u_\theta(x, y, t) = \pi(x, y)$, and the **efficient** type that cares about the total material payoffs in a matched pair, i.e. $u_\theta(x, y, t) = \pi(x, y) + \pi(y, x)$.

For our main analysis, we shall only consider a population with two different preference types θ and τ, where $\theta, \tau \in \Theta$.[11] A proportion $1 - \varepsilon$ of the agents carry θ and the remaining agents carry τ, where $\varepsilon \in (0, 1)$. We refer to the tuple $(\theta, \tau, \varepsilon)$ as a **population state**. Departing slightly from the existing literature, we do not place any restrictions on the magnitude of ε, allowing for a flexible interpretation of population states. When ε is close to 1, we can view θ as the invading minority in a population dominated by another type. Conversely, when ε is close to 0, θ can be seen as the incumbent type being invaded by a mutant type.

3 Preference Evolution with Complete Information

In this section, we assume that each agent observes the preference types of all other agents. Hence, when two agents are matched, they play Γ with complete information.

Fix a population state $(\theta, \tau, \varepsilon)$. For each type $t \in \{\theta, \tau\}$, we let $\mu_t \in \Delta(\{\theta, \tau\})$ be a probability distribution over types in the population that describes how type-t agents are matched. A **matching profile** is a vector $\mu = (\mu_\theta, \mu_\tau)$ that satisfies the following consistency condition:

$$(1 - \varepsilon)\mu_\theta[\tau] = \varepsilon\mu_\tau[\theta].$$

The condition above requires that the total mass of type-θ agents matched with type-τ agents is equal to that of type-τ agents matched with type-θ agents.[12]

Fixing a matching profile μ, for any $t, t' \in \{\theta, \tau\}$, let $s_{t,t'} \in \Delta(\mathcal{X}^2)$ describe the distribution of strategy pairs played across matches between type-t and type-t' agents, where the first component in $\mathcal{X}^2$ represents the strategy played by type t. An associated **strategy profile** $S = (s_{t,t'})$ is a vector of distributions of strategy pairs that satisfy the following exchangeability condition: Let $\rho : \mathcal{X}^2 \to \mathcal{X}^2$ be a mapping that switches the order of strategies, i.e. $\rho(x, y) = (y, x)$; then we have $s_{t,t'}[E] = s_{t',t}[\rho(E)]$ for any measurable set $E \subseteq \mathcal{X}^2$. When this condition is satisfied for $t' = t$, we say $s_{t,t}$ is exchangeable.

[11] In Section 5, we discuss how our results extend to the more general case of polymorphic populations.

[12] Although we take a distributional approach in defining the matching profile, there is no randomness in how agents are matched. Given that the population consists of finitely many types, one can explicitly describe the matching pattern through a deterministic mapping that generates μ.

We call the combination of a matching profile and an associated strategy profile (μ, S) an **outcome**.

Remark 1. In our model, when two agents of the same type are matched, they are allowed to play different strategies.[13] This is more general than the standard assumption in the literature on preference evolution, where the strategy pair has to be symmetric when agents of the same type are matched.[14] Importantly, as shown in Example 1, for some underlying games, there cannot be a stable outcome in which agents of the same type play the same strategy. Hence, the possibility of asymmetry is critical for our analysis.

3.1 Stable Matching

Given a population state $(\theta, \tau, \varepsilon)$, our next goal is to identify the outcomes (μ, S) that can be deemed *stable*. The requirement of stability has two layers. First, holding the matching profile μ fixed, agents do not want to change their strategies as specified by S. In other words, the strategy profile should constitute a Nash equilibrium. Second, given the utilities agents derive in an outcome, there should not exist agents who want to form a pairwise deviation and mutually benefit from rematching. To formalize this idea, we extend the notion of Nash stability by Garrido-Lucero and Laraki (2021) to a continuous population; see also Jackson and Watts (2010). We assume that agents cannot commit to a strategy via forces such as binding contracts or the possibility of future punishment in repeated interactions. This assumption restricts the set of pairwise deviations that are viable.

We now formally define these two layers of stability. A strategy profile S associated with μ is a **Nash equilibrium profile** if it satisfies the following: For $t, t' \in \{\theta, \tau\}$, if $\mu_t[t'] > 0$ and $(x^*, y^*) \in \operatorname{supp}(s_{t,t'})$, we have $x^* \in \arg\max_{x \in \mathcal{X}} u_t(x, y^*, t')$ and $y^* \in \arg\max_{y \in \mathcal{X}} u_{t'}(y, x^*, t)$. That is, every matched pair is playing a Nash equilibrium under S.

Definition 1. Fix an outcome (μ, S). We say there is a **blocking pair** if there exist types $t, t' \in \{\theta, \tau\}$ and a strategy pair $(\widehat{x}, \widehat{y}) \in \mathcal{X}^2$ such that for some types $\bar{t}$, $\bar{t}'$ and strategy pairs (x', y'), (x'', y''), we have

(i) $\mu_t[\bar{t}] > 0$, $\mu_{t'}[\bar{t}'] > 0$, $(x', y') \in \operatorname{supp}(s_{t,\bar{t}})$, and $(x'', y'') \in \operatorname{supp}(s_{t',\bar{t}'})$;

(ii) $\widehat{x} \in \arg\max_{x \in \mathcal{X}} u_t(x, \widehat{y}, t')$ and $\widehat{y} \in \arg\max_{y \in \mathcal{X}} u_{t'}(y, \widehat{x}, t)$;

[13] Note that $s_{t,t}$ being exchangeable does not imply symmetric play. For example, if $s_{t,t}$ is such that $s_{t,t}[(x, y)] = s_{t,t}[(y, x)] = \frac{1}{2}$ where $x \neq y$, then type-t agents play an asymmetric strategy pair (x, y) across all same-type matches.

[14] The rationale for the standard assumption in the literature is as follows. Since the underlying two-person game Γ is simultaneous and the matching process is exogenous, there is no opportunity for the agents to condition their strategies on their matched partners' strategies, but only their types. Therefore, given that Γ is symmetric, if $x_{\theta,\theta} \in X$ denotes the strategy chosen by a θ-type agent against another θ-type agent, then for a pair of θ-type agents, they both necessarily play $x_{\theta,\theta}$.

(iii) $u_t(\widehat{x}, \widehat{y}, t') > u_t(x', y', \bar{t})$ and $u_{t'}(\widehat{y}, \widehat{x}, t) > u_{t'}(x'', y'', \bar{t}')$.

This notion of blocking pair for aggregate matching is analogous to the one in Echenique et al. (2013), which serves as a natural generalization of the blocking concept proposed by Gale and Shapley (1962) to continuous populations. Condition (i) says the agents participating in a blocking pair must have positive mass. Condition (ii) requires the deviating agents agree on a Nash equilibrium so that their strategies are mutual best responses; that is, the deviation is *credible*. Finally, condition (iii) means the deviating agents strictly prefer to rematch, which means the proposed strategy pair is indeed *profitable*.

Definition 1 (iii) requires strict incentives to rematch for both parties. Alternatively, this definition can be relaxed to allow only one side of the blocking pair to have a strict incentive. However, this weaker definition may lead to non-existence of stable outcomes. Additionally, our main results remain unchanged even if this alternative definition is adopted. See Online Appendix O.1 for a thorough discussion.

As we argued above, for an outcome to be stable, the status quo should constitute a Nash equilibrium for each pair of matched agents, and there should be no profitable blocking pair under the outcome. Therefore, we have the following definition of Nash stability.

Definition 2. An outcome (μ, S) is **Nash stable** if it satisfies:

(i) S is a Nash equilibrium profile **(internal stability)**;

(ii) There is no blocking pair **(external stability)**.

The notion of Nash stability describes the outcomes that, once reached, do not induce further strategic or coalitional adjustments. As a direct generalization of pairwise stability, agents are assumed to be shortsighted in the sense that they only compare one-shot utilities in a pairwise deviation. Since there is no predetermined sides in our model, the matching problem resembles a "roommate problem," which does not guarantee a stable outcome with finitely many agents (Gale and Shapley, 1962). However, Nash stable outcomes always exist in our model with a continuum of agents, a result we prove in Proposition 8 for the more general setting that allows for polymorphism. Next, we use an example to illustrate Definition 2.

Example 1. Consider a population state $(\theta, \tau, \varepsilon)$. The utility functions of the two types are described by the following three scenarios of strategic interactions:

When two type-θ agents are matched, there are three Nash equilibria: (A, B), (B, A), and $(\frac{1}{2}A + \frac{1}{2}B, \frac{1}{2}A + \frac{1}{2}B)$. Since A is the dominant strategy for the type-τ agents, the only Nash equilibrium between two type-τ agents is (A, A). When a type-θ agent is matched with a type-τ agent, the only Nash equilibrium is (B, A). We now argue that any Nash

type θ \ type θ	A	B
A	$0,0$	$3,3$
B	$3,3$	$0,0$

type θ \ type τ	A	B
A	$0,4$	$3,3$
B	$3,3$	$0,0$

type τ \ type τ	A	B
A	$4,4$	$3,3$
B	$3,3$	$0,0$

stable outcome (μ, S) must satisfy $\mu_\theta[\theta] = \mu_\tau[\tau] = 1$, $s_{\theta,\theta}[(A,B)] = s_{\theta,\theta}[(B,A)] = \frac{1}{2}$, and $s_{\tau,\tau}[(A,A)] = 1$.[15] That is, type-θ agents are only matched with type-θ agents, while type-τ agents are only matched with type-τ agents; each pair of type-θ agents play the strategy pair (A,B), and each pair of type-τ agents play the strategy pair (A,A).

To see this, suppose the contrary. There are two cases to consider:

1) $\mu_\theta[\tau] > 0$. In this case, any θ-τ pair must play the unique Nash equilibrium (B,A), where the type-τ agent obtains a utility of 3. However, these type-τ agents who are matched with type-θ agents can form a Nash blocking pair and benefit from playing their dominant strategy equilibrium (A,A), violating external stability.

2) $\mu_\theta[\theta] = \mu_\tau[\tau] = 1$ but $s_{\theta,\theta}[(\frac{1}{2}A + \frac{1}{2}B, \frac{1}{2}A + \frac{1}{2}B)] > 0$. In this case, a positive mass of type-θ agents derive a utility of $\frac{3}{2}$. However, they can form a blocking pair and play a pure strategy Nash equilibrium (A,B), where both sides in the rematch obtain a utility of $3 > \frac{3}{2}$.

Finally, we verify that (μ, S) is indeed Nash stable. First note that each matched pair is playing a Nash equilibrium. Thus, the outcome is internally Nash stable. For external Nash stability, observe that all type-τ and type-θ agents already obtain their highest possible utilities, which means they can never be made better off in a deviation. Therefore, no Nash blocking pair exists.

$\diamondsuit$

A few remarks are in order. First, in Example 1, a Nash stable outcome must be *asymmetric* in the sense that half of the type-θ agents play strategy A, while the other half play B. This demonstrates that coordination on asymmetric strategy pair is a possible and natural outcome in our model. The driving force behind this is our consideration of endogenous partner choice. In particular, agents can engage in communication while negotiating a credible and profitable pairwise deviation and can maintain the asymmetric

[15] The Nash stable outcome is unique in a generic sense because $s_{\theta,\tau}$ and $s_{\tau,\theta}$ can be arbitrarily specified for a set of pairs with measure zero, and they do not have implications on the Nash stability of (μ, S).

play with a particular partner. This is in sharp contrast to the existing literature which only considers symmetric equilibria. Moreover, observe that in Example 1, there are three Nash equilibrium strategy pairs between type-θ agents, but only two of them, (A, B) and (B, A), are played in a Nash stable outcome. Therefore, stable matching has implications for *equilibrium selection* in our setting. The following definition captures this equilibrium selection effect.[16]

Definition 3. For $t \in \Theta$, let $NE_t \subseteq \mathcal{X}^2$ denote the set of Nash equilibria between two type-t agents. Define the set of **loser-best Nash equilibria** between type-t agents as

$$NE_t^{lb} = \arg\max_{(x,y)\in NE_t} \min\{u_t(x, y, t), u_t(y, x, t)\}.$$

Note that the set NE_t^{lb} is nonempty because $u_t(x, y, t)$ is continuous in (x, y) and the set NE_t is compact. We now make an immediate observation. All the proofs for the results presented in this paper are relegated to the Appendix.

Lemma 1. *In a population state $(\theta, \tau, \varepsilon)$, suppose there exists a Nash stable outcome (μ, S) with $\mu_t[t] > 0$ for $t \in \{\theta, \tau\}$. Then $s'_{t,t} \in S'$ for some Nash stable outcome (μ, S') if and only if $s'_{t,t}$ is exchangeable and $s'_{t,t}[NE_t^{lb}] = 1$.*

This lemma captures the equilibrium selection effect we observed in Example 1. In particular, the mixed strategy Nash equilibrium between type-θ agents is not loser-best, and thus it cannot be played in any Nash stable outcome.

3.2 Evolutionary Stability

Given a Nash stable outcome (μ, S) in population state $(\theta, \tau, \varepsilon)$, the average material payoffs for type-θ and type-τ agents are given by

$$G_\theta(\mu, S) = \sum_{t\in\{\theta,\tau\}} \mu_\theta[t] \int_{(x,y)\in\mathcal{X}^2} \pi(x, y)\, ds_{\theta,t},$$
$$G_\tau(\mu, S) = \sum_{t\in\{\theta,\tau\}} \mu_\tau[t] \int_{(x,y)\in\mathcal{X}^2} \pi(x, y)\, ds_{\tau,t}\,.$$

We now define the notion of evolutionary stability as follows.

[16] This effect shares a similar spirit with the one analyzed in Jackson and Watts (2010). In both our and their settings, stability puts restrictions on the outcome and therefore refines the set of Nash equilibria that can arise.

Definition 4. A preference type $\theta \in \Theta$ is **evolutionarily stable against** another type $\tau \in \Theta$ if for every $\varepsilon \in (0,1)$, in population state $(\theta, \tau, \varepsilon)$, $G_\theta(\mu, S) \geqslant G_\tau(\mu, S)$ for all Nash stable outcomes (μ, S) while the inequality is strict for some Nash stable outcome. A preference type θ is **evolutionarily unstable** if there exists another type τ that is evolutionarily stable against θ.

Remark 2. Our definition of evolutionary stability is neither stronger nor weaker than the one in Alger and Weibull (2013). First, they require strict inequality for all Nash equilibria under their exogenous matching process, while we only require it for some Nash stable outcome. On the other hand, and more importantly, their notion of evolutionary stability is defined in a local sense, while ours is a *global* one as the inequality should hold regardless of the proportion ε of type τ. Accordingly, we require that the evolutionarily stable type θ not only resists invasion when it is the incumbent type (i.e., ε is close to 0) but also has the ability to invade the population when it is the mutant (i.e., ε is close to 1).

We only define evolutionary stability against a particular type τ because requiring the condition to hold against all possible types would be too stringent given that Θ is rich: For example, if another type τ never wants to match with θ and behave just like θ among themselves, then the average material payoffs would be the same across the two types. Next, we introduce a related notion called neutral stability:

Definition 5. A preference type $\theta \in \Theta$ is **neutrally stable** if for every $\tau \in \Theta$ and $\varepsilon \in (0,1)$, in population state $(\theta, \tau, \varepsilon)$, $G_\theta(\mu, S) \geqslant G_\tau(\mu, S)$ for all Nash stable outcomes (μ, S).

While neutral stability only requires a weak inequality for all Nash stable outcomes, the inequality should hold for all types $\tau \in \Theta$. It is closer in spirit to the notion of stability considered in Dekel et al. (2007) (except that they consider a local notion). By definition, if a type evolutionarily unstable, it is not neutrally stable. Given the definitions of evolutionary (un)stability, we now proceed with the analysis.

As a standard terminology, we say a strategy pair $(\widetilde{x}, \widetilde{y})$ is **efficient** if[17]

$$(\widetilde{x}, \widetilde{y}) \in \arg\max_{(x,y)\in\mathcal{X}^2} \pi(x,y) + \pi(y,x),$$

and let M denote the total material payoff generated by an efficient strategy pair. A strategy pair is **inefficient** if it is not efficient. Efficiency plays an important role in the subsequent analysis because preference evolution is driven by material payoff success.

[17]Note that in previous literature, since agents of same type have to play the same strategy, the consideration of efficiency is restricted to symmetric strategy profiles. See for example, Dekel et al. (2007).

Definition 6. We say θ **exhibits same-type inefficiency** if there exists a loser-best Nash equilibrium between type-θ agents that is inefficient.

Note that this definition also imposes an implicit but weak assumption on the material game Γ: It must have an inefficient strategy pair. One main message of this paper is that efficient play is the only possible outcome that can sustain in the long run. We next identify two kinds of plastic preferences that can ensure efficient play with complete information.

Definition 7. For $\alpha > 0$, a preference type θ is called the α**-homophilic efficient** type if the corresponding utility function takes the form

$$u_\theta(x, y, t) = \pi(x, y) + \pi(y, x) + \alpha \cdot \mathbb{1}_{\{t=\theta\}}. \tag{1}$$

Any preference type in this class is called **homophilic efficient**.

In the network literature, the tendency of people to interact with similar people is referred to as homophily (see Jackson (2014)). We model this tendency on a preference level: A α-homophilic efficient agent has a natural inclination to interact with another α-homophilic efficient agent because she can derive an extra utility of α.

Definition 8. A preference type θ is called the **parochial efficient** type if the corresponding utility function takes the form

$$u_\theta(x, y, t) = [\pi(x, y) + \pi(y, x)] \cdot \mathbb{1}_{\{t=\theta\}}. \tag{2}$$

A parochial efficient agent has a strong tendency to be associated with another parochial efficient agent because it is the only possibility that she can derive a positive utility. Hence, one can consider parochialism as a strong form of homophily. Newton (2017) also considers parochialism in preference evolution. He defines parochialism on the matching level, meaning that the parochial agents are only matched with one another. On the contrary, we define parochialism on the preference level, and how parochial agents are matched is determined by stable matching.

Our first result shows that a preference for efficiency with any level of homophily or with parochialism is likely to be the type that prevails in the long run.

Proposition 1. *The homophilic efficient and parochial efficient types are neutrally stable. Moreover, they are evolutionarily stable against any type that exhibits same-type inefficiency.*

Proposition 1 shows that efficiency is the driving force for evolutionary selection of preferences under stable matching, which is similar to what Dekel et al. (2007) demonstrate for preference evolution under random matching (although our definition of efficiency is

more general and allows for asymmetry). However, our mechanism supporting efficiency is grounded in endogenous assortative matching, in contrast to their reliance on the "secret handshake" idea introduced by Robson (1990). In our model, the evolutionary stability relies on two features of the behavior: efficient play and a preference for matching with the same type. Take a homophilic preference type as an example. The homophilic component of the utility function ensures that, although these agents aim to play efficiently with their matched partners, they have a strict incentive to do so with others who have the same preference type. Such a self-match incentive induces positive assortative matching in the population, which ensures that the homophilic-efficient agents will not be taken advantage of by other type agents and play efficiently among themselves exclusively.

Example 2. Consider a material game where each player has three pure strategies. The material payoffs are given by the payoff matrix below.

	A	B	C
A	$0,0$	$0,0$	$2,8$
B	$0,0$	$3,3$	$4,0$
C	$8,2$	$0,4$	$0,0$

Let θ be a homophilic or parochial efficient type. In this game, (A,C) and (C,A) are the only efficient strategy pairs. It is important to observe that they are indeed Nash equilibria for two type-θ agents because, as long as the partner is playing C (or A), the strategy A (or C, respectively) maximizes the total material payoff. Note that the strategy pair (B,B) is another Nash equilibrium for type-θ agents, but it is not loser-best. By Proposition 1, type θ can prevail in evolution because it is able to ensure assortative matching and coordination on the efficient strategy pairs (A,C) and (C,A). ♢

Example 2 illustrates the stark difference between models based on random matching and our model based on stable matching. Dekel et al. (2007) show that incumbents playing a symmetrically efficient strategy pair (a symmetric strategy pair that maximizes the total material payoff among all symmetric strategy pairs) is a necessary condition for evolutionary stability, and symmetric efficiency together with strict Nash is a sufficient condition. In Example 2, (B,B) is symmetrically efficient and a strict Nash equilibrium (of the material game). However, incumbents playing it cannot withstand the invasion of mutants who are able to force positive assortative matching and play the more efficient asymmetric strategy pairs (A,C) and (C,A).

While we find that efficiency combined with homophily or parochialism is most natural and serves as a sufficient condition for evolutionary stability, we do not claim that these are the only preference types can be neutrally stable and evolutionarily stable against any other

type that exhibits same-type inefficiency.[18] We next argue that both efficient play and a preference for matching with the same type are necessary for resistance against mutation.

Proposition 2. (i) *If θ exhibits same-type inefficiency, then θ is evolutionarily unstable;*

(ii) *If $\pi(\widetilde{x},\widetilde{y}) \neq \pi(\widetilde{y},\widetilde{x})$ for every efficient strategy pair $(\widetilde{x},\widetilde{y})$, and $u_\theta(x,y,t)$ is constant in t, then θ is evolutionarily unstable.*

As an implication of this result, if θ is neutrally stable, then it cannot exhibit same-type inefficiency; moreover, it cannot be indifferent about the opponent's type when efficient outcomes are asymmetric. Part (i) of Proposition 2 is a direct corollary of Proposition 1. The intuition behind part (ii) is as follows. When type-θ agents do not have plastic preferences, they may not be able to induce positive assortative matching, meaning that they have a chance to be matched with other types in the population. Then, if type-τ agents play an efficient strategy pair among themselves, and utilize the asymmetry of an efficient strategy pair in the cross-type matches by committing to the advantageous strategy, they would obtain a higher average material payoff than type-θ agents.[19] The asymmetry of efficient outcomes is inherent in a wide range of strategic interactions, as efficiency is typically enhanced by specialization in behavior due to complementarity, which can, in turn, lead to unbalanced material payoffs.

Part (ii) of Proposition 2 includes the rich set of preference types studied in the literature of preference evolution that do not exhibit plasticity. Typical examples include preferences that represent spite, selfishness, or altruism, i.e. $u_\theta(x,y,t) = \pi(x,y) + \alpha\pi(y,x)$ with $\alpha < 0$, $\alpha = 0$, or $\alpha > 0$; and homo-moralis, i.e. $u_\theta(x,y,t) = (1-\alpha)\pi(x,y) + \alpha\pi(x,x)$, with $\alpha \in [0,1]$. It demonstrates that, with endogenous partner choice, these non-plastic preference types cannot prevail in games without symmetric efficient strategy pairs.

3.3 Selfishness and Nash Equilibria

Selfishness has been proven to be not favored by preference evolution under random matching with complete information since the work of Güth and Yaari (1992) and Güth (1995), because a population of selfish agents can be destabilized by "secret handshake" of the mutants. Under stable matching, can selfishness be stable if combined with some form of plasticity?

[18] For instance, a preference type can prevail as long as playing an efficient strategy pair with others of the same type yields the highest possible utility, while the remaining details of the utility function can be specified arbitrarily.

[19] Commitment works in our model because we allow that an agent's utility exhibits plasticity. For example, assume that type τ's utility function is given by $u_\tau(x,y,\tau) = \pi(x,y) + \pi(y,x)$ and $u_\tau(x,y,\theta) = \alpha \cdot \mathbb{1}_{\{x=x'\}}$ for some $\alpha > 0$. In this case, type-τ agents would play an efficient strategy pair among themselves but are "committed" to playing x' against type-θ opponents.

Definition 9. With $\alpha > 0$, a preference type θ is called the **α-homophilic selfish** type if the corresponding utility function takes the form

$$u_\theta(x, y, t) = \pi(x, y) + \alpha \cdot \mathbb{1}_{\{t=\theta\}}. \tag{3}$$

Any preference type in this class is called **homophilic selfish**.

Definition 10. A preference type θ is called the **parochial selfish** type if the corresponding utility function takes the form

$$u_\theta(x, y, t) = \pi(x, y) \cdot \mathbb{1}_{\{t=\theta\}}. \tag{4}$$

Write NE_π for the set of Nash equilibria in the material game and NE_π^{lb} for the set of loser-best Nash equilibria between selfish agents.

Proposition 3. *Suppose all strategy pairs in NE_π^{lb} are efficient.*

(i) *If α is sufficiently large, then the α-homophilic selfish type is neutrally stable and evolutionarily stable against any type that exhibits same-type inefficiency;*

(ii) *The parochial selfish type is neutrally stable and evolutionarily stable against any type that exhibits same-type inefficiency.*

In general, if some strategy pair in NE_π^{lb} is inefficient, then any homophilic selfish or parochial selfish type exhibits same-type inefficiency, which means it is evolutionarily unstable by Proposition 2.

Example 3 (Example 2 revisited)**.** Consider again the material game in Example 2. Suppose θ is a homophilic or parochial selfish type. The unique (loser-best) Nash equilibrium (B, B) between two type-θ agents is inefficient. Let τ be the parochial efficient type. For any population state $(\theta, \tau, \varepsilon)$, a Nash stable outcome must be perfectly assortative and satisfy $s_{\theta,\theta}[(B, B)] = 1$. Because the parochial efficient type can coordinate on the efficient strategy pairs (A, C) and (C, A), type θ fares strictly worse than type τ in terms of average material payoffs. In other words, type θ is evolutionarily unstable. ♢

Proposition 4 considers the special case that a symmetric strategy pair happens to be both efficient and a Nash equilibrium, where selfish types become evolutionarily stable.

Proposition 4. *Suppose there exists a symmetric strategy pair $(\widetilde{x}, \widetilde{x})$ that is an efficient Nash equilibrium of the material game. Then the homophilic selfish and parochial selfish types are neutrally stable. Moreover, they are evolutionarily stable against any type that exhibits same-type inefficiency.*

4 Preference Evolution with Incomplete Information

In this section, we turn our attention to the case of incomplete information. Suppose that in the population, every agent knows her own preference type, but may not observe the types of other agents.[20] Fixing a population state $(\theta, \tau, \varepsilon)$, we need to generalize the notion of matching profile to encompass incomplete information.

A **matching profile** (with incomplete information) is a tuple (Λ, p, q, μ). The first component Λ is a finite set of **labels** that are publicly observable. The population is further described by a probability distribution with full support $p \in \Delta(\Lambda)$ over the set of labels. Each label $\lambda \in \Lambda$ is associated with a probability distribution $q_\lambda \in \Delta(\{\theta, \tau\})$ over preference types. We assume $q_\lambda \neq q_{\lambda'}$ whenever $\lambda \neq \lambda'$, reflecting that different labels convey distinct information, and write $q = (q_\lambda)_{\lambda \in \Lambda}$. The pair (p, q) should satisfy the following marginal condition:

$$\sum_{\lambda \in \Lambda} p[\lambda] q_\lambda[\theta] = 1 - \varepsilon.$$

In words, the masses of type-θ agents with different labels should sum up to their total mass in the population. In the following analysis, we will refer to a type-θ agent with label λ simply as a type-θ_λ agent.[21] Analogous to the case of complete information, for any $\lambda \in \Lambda$, we let $\mu_\lambda \in \Delta(\Lambda)$ be a probability distribution over labels that describes how label-λ agents are matched. The last component of a matching profile is then a vector $\mu = (\mu_\lambda)_{\lambda \in \Lambda}$ that satisfies the consistency condition below

$$p[\lambda]\mu_\lambda[\lambda'] = p[\lambda']\mu_{\lambda'}[\lambda], \quad \text{for all } \lambda, \lambda' \in \Lambda.$$

Given a matching profile (Λ, p, q, μ), for any $\lambda, \lambda' \in \Lambda$, we let $s_{\lambda,\lambda'} \in \Delta(\mathcal{X}^2)$ describe the distribution of strategy pairs played across matches between label-λ and label-λ' agents. An associated **strategy profile** $S = (s_{\lambda,\lambda'})$ is a vector of distributions of strategy pairs that satisfy the exchangeability condition as in the case of complete information. Moreover, for any $\lambda \in \Lambda$, we assume the strategy distribution $s_{\lambda,\lambda'}$ is independent of the informational content of labels, q_λ; that is, belief updating from q_λ is constant across realizations of the strategy pair from $s_{\lambda,\lambda'}$. This is because strategies are assumed to be observable, so all information

[20] Recall that our model allows plasticity, meaning an agent's utility function can depend on the preference type of her matched partner. One interpretation is that an agent may value certain characteristics of her opponent, which are perfectly correlated with preferences. When these characteristics are readily observable (e.g. physical appearance), a model with complete information suffices. However, if these characteristics are intrinsically hidden (e.g. empathy or sense of responsibility), we must employ a model that accounts for incomplete information. Relaxing the assumption of perfect correlation between characteristics and preferences is conceptually straightforward but beyond the scope of this paper. See a discussion in the concluding section.

[21] Note that "labels" are purely informational and do not affect utilities.

inferred from these observations should be already encoded in the labels.

As before, the combination of a matching profile and an associated strategy profile (Λ, p, q, μ, S) is called an **outcome** (with incomplete information).

Remark 3. The informational component (Λ, p, q) is a part of the outcome, rendering information *endogenous*, as is standard in the literature of matching with incomplete information (Liu et al., 2014; Liu, 2020; Chen and Hu, 2023; Wang, 2022, etc.). In this paper, we make no exogenous informational assumptions and consider the set of all outcomes with publicly observed labels. Naturally, certain assumptions can be incorporated by, for instance, introducing a commonly understood signal structure.[22] All our results remain valid under such exogenous informational assumptions, provided they do not fully disclose all information—that is, as long as some level of incomplete information persists.

4.1 Stable Matching with Incomplete Information

To simplify notation, for $t \in \{\theta, \tau\}$ and $\lambda \in \Lambda$, we write

$$u_t(x, y, \lambda) = q_\lambda[\theta] u_t(x, y, \theta) + q_\lambda[\tau] u_t(x, y, \tau),$$

which is the expected utility of a type-t agent when playing (x, y) with a label-λ partner. Fixing a population state $(\theta, \tau, \varepsilon)$, a strategy profile S associated with (Λ, p, q, μ) is a **Bayes-Nash equilibrium profile** if the following condition is satisfied:[23] For $\lambda, \lambda' \in \Lambda$, if $\mu_\lambda[\lambda'] > 0$ and $(x^*, y^*) \in \operatorname{supp}(s_{\lambda,\lambda'})$, we have $x^* \in \arg\max_{x \in \mathcal{X}} u_t(x, y^*, \lambda')$ for each $t \in \operatorname{supp}(q_\lambda)$ and $y^* \in \arg\max_{y \in \mathcal{X}} u_{t'}(y, x^*, \lambda)$ for each $t' \in \operatorname{supp}(q_{\lambda'})$. In words, every agent playing against a label-λ partner plays a best response with the belief that the partner is of type-θ_λ with probability $q_\lambda[\theta]$ and of type-τ_λ with the complementary probability. When incomplete information is absent, i.e. there are only two labels each associated with a degenerate distribution, the definition above reduces to the notion of Nash equilibrium profile defined in Section 3.1.

Because agents can only recognize the labels but not the preference types of potential partners, we need another definition to properly define pairwise deviations under incomplete

[22] To illustrate, consider a simple example. Suppose agents' types in a population state $(\theta, \tau, \varepsilon)$ can be revealed before matching via the following signal structure (ξ_θ, ξ_τ): For a type-t agent, $t \in \{\theta, \tau\}$, a signal that perfectly reveals her preferences is generated and publicly observed with probability $\xi_t > 0$, while no signal is observed with complimentary probability $1 - \xi_t$. Under this signal structure, the outcomes (Λ, p, q, μ, S) must satisfy the following conditions: There exist $\lambda, \lambda' \in \Lambda$ such that $q_\lambda[\theta] = 1$, $q_{\lambda'}[\tau] = 1$, $p[\lambda] \geqslant (1 - \varepsilon)\xi_\theta$, and $p[\lambda'] \geqslant \varepsilon\xi_\tau$. Intuitively, these conditions require that the any outcome is at least as informative as the initial signal structure.

[23] In our setting, agents observe their partners' behavior but do not infer additional information from these observations. Therefore, our notion of Bayes-Nash equilibrium profile shares similarities with the rational expectations equilibrium studied in Koh (2023).

information.

Definition 11. A pair $(D_\lambda, \boldsymbol{x}_\lambda)$ is a **deviation plan** for a label-λ agent if (i) D_λ is a nonempty subset of $\text{supp}(q_\lambda)$, and (ii) $\boldsymbol{x}_\lambda : D_\lambda \to \mathcal{X}$.

In words, D_λ is the set of preference types that have label λ poised to participate in a pairwise deviation, while $\boldsymbol{x}_\lambda$ is a mapping that specifies a strategy played by each deviating type.

Definition 12. Fix an outcome with incomplete information (Λ, p, q, μ, S). We say an **incomplete information blocking pair** exists if there exist types $t, t' \in \{\theta, \tau\}$ and labels $\lambda, \lambda' \in \Lambda$ with $q_\lambda[t] > 0$ and $q_{\lambda'}[t'] > 0$ such that for some labels $\bar{\lambda}$, $\bar{\lambda}'$ and strategy pairs (x', y') and (x'', y'')

(i) $\mu_\lambda[\bar{\lambda}] > 0$, $\mu_{\lambda'}[\bar{\lambda}'] > 0$, $(x', y') \in \text{supp}(s_{\lambda,\bar{\lambda}})$, and $(x'', y'') \in \text{supp}(s_{\lambda',\bar{\lambda}'})$;

Moreover, there exists a strategy pair $(\widehat{x}, \widehat{y})$ such that for any deviation plans $(D_\lambda, \boldsymbol{x}_\lambda)$ and $(D_{\lambda'}, \boldsymbol{y}_{\lambda'})$ with $t \in D_\lambda$, $t' \in D_{\lambda'}$, $\boldsymbol{x}_\lambda(t) = \widehat{x}$, and $\boldsymbol{y}_{\lambda'}(t') = \widehat{y}$, we have

(ii) $\widehat{x} \in \arg\max_{x \in \mathcal{X}} \mathbb{E}_{q_{\lambda'}}[u_t(x, \boldsymbol{y}_{\lambda'}(\cdot), \cdot) \,|\, D_{\lambda'}]$ and $\widehat{y} \in \arg\max_{y \in \mathcal{X}} \mathbb{E}_{q_\lambda}[u_{t'}(y, \boldsymbol{x}_\lambda(\cdot), \cdot) \,|\, D_\lambda]$;

(iii) $\mathbb{E}_{q_{\lambda'}}[u_t(\widehat{x}, \boldsymbol{y}_{\lambda'}(\cdot), \cdot) \,|\, D_{\lambda'}] > u_t(x', y', \bar{\lambda})$ and $\mathbb{E}_{q_\lambda}[u_{t'}(\widehat{y}, \boldsymbol{x}_\lambda(\cdot), \cdot) \,|\, D_\lambda] > u_{t'}(x'', y'', \bar{\lambda}')$.

In the definition above, for an agent of type $t \in \{\theta, \tau\}$, her strategy $x \in \mathcal{X}$, and a deviation plan $(D_{\lambda'}, \boldsymbol{y}_{\lambda'})$ of the deviating partner with label λ', the conditional expected utility $\mathbb{E}_{q_{\lambda'}}[u_t(x, \boldsymbol{y}_{\lambda'}(\cdot), \cdot) \,|\, D_{\lambda'}]$ is evaluated using the probability distribution $q_{\lambda'}$ conditional on the subset of types $D_{\lambda'}$. If $D_{\lambda'}$ is a singleton, then the expectation is degenerate.

Definition 12 describes a situation in which a pair of agents, despite observing only each other's label, can still reach an agreement and carry out a mutually beneficial deviation. In particular, the deviating agents are of types t_λ and $t'_{\lambda'}$, which we call the "targeted" types. As long as the targeted type-$t'_{\lambda'}$ agents participate and play $\widehat{y}$, the deviating type-t_λ agent will play $\widehat{x}$ as a best response which strictly improves her utility, *regardless of* whether and how the non-targeted agents with label λ' participate in the deviation. The same reasoning applies to the deviating type-$t'_{\lambda'}$ agent. In other words, the incentives to deviate are conditional on the participation of the targeted partners in the blocking pair. When the targeted types are fully revealed by their labels, i.e. when both q_λ and $q_{\lambda'}$ are degenerate, the conditions in Definition 12 reduce to those in Definition 1 under complete information.

Remark 4. In Definition 12, a deviating agent believes in rationality of her targeted partners conditional on her own participation. One might further account for rationality of the non-targeted types as well. However, we adopt a more conservative approach for several

reasons. First, rational behavior in a deviation may reveal additional information,[24] potentially triggering the deviating partner to adjust her behavior from the original plan. Addressing this requires us to take a stance on how agents anticipate and respond to such possibilities; see a related discussion in Liu (2020). Our definition avoids this issue since the non-targeted agents' responses do not influence the decision-making of the targeted agents. Second, adopting a more stringent definition of blocking weakens the concept of stability, thereby strengthening the positive result in Proposition 6. Meanwhile, the negative result in Proposition 5 does not depend on any irrational behavior of the non-targeted type (see footnotes 28 and 31).

We use an example to illustrate the notion of incomplete information blocking pair.

Example 4. Consider the prisoners' dilemma material game as follows:

	A	B
A	$4,4$	$0,5$
B	$5,0$	$3,3$

Suppose type-θ agents have efficient preferences $u_\theta(x,y,t) = \pi(x,y)+\pi(y,x)$, while type-$\tau$ agents are selfish $u_\tau(x,y,t) = \pi(x,y)$. Consider a population state $(\theta,\tau,\varepsilon)$ and an outcome (Λ,p,q,μ,S) as follows. There is a half of each type in the population, i.e. $\varepsilon = \frac{1}{2}$. The matching profile (Λ,p,q,μ) satisfies $\Lambda = \{\lambda\}$, $p[\lambda] = 1$, $q_\lambda[\theta] = q_\lambda[\tau] = \frac{1}{2}$, and $\mu_\lambda[\lambda] = 1$. The strategy profile is $S = \{s_{\lambda,\lambda}\}$ with $s_{\lambda,\lambda}[(B,B)] = 1$. This outcome is depicted in Figure 1 below.

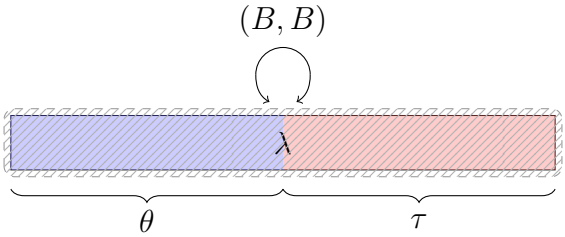

Figure 1: The matching profile in Example 4.

To see that an incomplete information blocking pair exists, consider two type-θ_λ agents who target each other and propose the efficient strategy pair (A,A). By symmetry, we only need to verify the following conditions: Facing any deviation plan $(D_\lambda, \boldsymbol{x}_\lambda)$ of the partner that satisfies $\theta \in D_\lambda$ and $\boldsymbol{x}_\lambda(\theta) = A$, a type-$\theta_\lambda$ agent plays A as a best response which strictly improves her utility. There are two cases to consider:

- $\tau \notin D_\lambda$. Here, A is a best response against A, which yields $8 > 6$;

[24] This occurs, for example, when the deviation plan $(D_\lambda, \boldsymbol{x}_\lambda)$ satisfies $D_\lambda = \{\theta,\tau\}$ and $\boldsymbol{x}_\lambda(\theta) \neq \boldsymbol{x}_\lambda(\tau)$.

- $\tau \in D_\lambda$ and $\boldsymbol{x}_\lambda(\tau) = \sigma A + (1-\sigma)B$. In this case, A is still a best response for a type-θ_λ agent because $\frac{1}{2} \cdot 8 + \frac{1}{2}(8\sigma + 5(1-\sigma)) > \frac{1}{2} \cdot 5 + \frac{1}{2}(5\sigma + 6(1-\sigma))$ for all $\sigma \in [0,1]$. Moreover, this makes her strictly better off, as $\frac{1}{2} \cdot 8 + \frac{1}{2}(8\sigma + 5(1-\sigma)) > 6$ for all $\sigma \in [0,1]$.

In summary, conditional on the fact that a type-θ_λ partner will participate in the deviation and play A, playing A is indeed a best response for a type-θ_λ agent and the deviation makes her strictly better off. This is true even if the non-targeted type-τ_λ agents join the deviation and play arbitrarily. ♢

We extend the notion of stable outcome to the case of incomplete information.

Definition 13. an outcome with incomplete information (Λ, p, q, μ, S) is **Bayes-Nash stable** if it satisfies:

(i) S is a Bayes-Nash equilibrium profile **(internal stability)**;

(ii) There is no incomplete information blocking pair **(external stability)**.

When $\Lambda = \{\lambda, \lambda'\}$, $p[\lambda] = 1-\varepsilon$, $p[\lambda'] = \varepsilon$, and $q_\lambda[\theta] = q_{\lambda'}[\tau] = 1$, Bayes-Nash stability reduces to Nash stability. Thus, the existence of a Bayes-Nash stable outcome is guaranteed. This existence argument is analogous to the one in the recent literature of matching with incomplete information (see, for example, Liu et al., 2014). Naturally, some Bayes-Nash stable outcomes may fail to satisfy Nash stability if preferences were fully observable. Since information is endogenous, the extend to which agents' preferences are revealed in a Bayes-Nash stable outcome depends on their preferences and behaviors in the game.

4.2 Evolutionary Stability with Incomplete Information

Given a Bayes-Nash stable outcome (Λ, p, q, μ, S) in population state $(\theta, \tau, \varepsilon)$, the average material payoffs for agents of type θ and type τ are given by

$$G_\theta(\Lambda, p, q, \mu, S) = \sum_{\lambda \in \Lambda} \frac{p_\lambda q_\lambda[\theta]}{1-\varepsilon} \left\{ \sum_{\lambda' \in \Lambda} \mu_\lambda[\lambda'] \int_{(x,y) \in \mathcal{X}^2} \pi(x,y)\, ds_{\lambda,\lambda'} \right\},$$

$$G_\tau(\Lambda, p, q, \mu, S) = \sum_{\lambda \in \Lambda} \frac{p_\lambda q_\lambda[\tau]}{\varepsilon} \left\{ \sum_{\lambda' \in \Lambda} \mu_\lambda[\lambda'] \int_{(x,y) \in \mathcal{X}^2} \pi(x,y)\, ds_{\lambda,\lambda'} \right\}.$$

Our notions of evolutionary stability and unstability can be naturally extended to incorporate incomplete information by replacing "Nash stable outcomes (μ, S)" with the more general "Bayes-Nash stable outcomes (Λ, p, q, μ, S)" in Definitions 4 and 5.

The reason that homophilic efficient types are evolutionarily stable under complete information (Proposition 1) is that they can always induce assortative matching and efficient play among themselves. The following example shows that the sorting mechanism no longer works under incomplete information.

Example 5. Consider a material game where each player has two strategies. The material payoffs are given in the following table:

	A	B
A	$0, 0$	$1, 3$
B	$3, 1$	$0, 0$

Let θ denote the α-homophilic efficient type with $\alpha > 0$. Consider a type τ that is selfish when playing with her own type, but has a dominant strategy B otherwise:

$$u_\tau(x, y, t) = \begin{cases} \pi(x, y) & \text{if } t = \tau, \\ 4 \cdot \mathbb{1}_{\{x=B\}} & \text{if } t \neq \tau. \end{cases}$$

Now consider a population state $(\theta, \tau, \varepsilon)$ and an outcome (Λ, p, q, μ, S) as follows. The proportion of type-τ agents satisfies $\varepsilon \geqslant \frac{2+\alpha}{4+\alpha}$. The matching profile (Λ, p, q, μ) satisfies $\Lambda = \{\lambda, \lambda'\}$, $p[\lambda] = p[\lambda'] = \frac{1}{2}$, $q_\lambda[\theta] = 2(1-\varepsilon) \leqslant \frac{4}{4+\alpha}$, $q_{\lambda'}[\tau] = 1$, and $\mu_\lambda[\lambda'] = \mu_{\lambda'}[\lambda] = 1$. The strategy profile is $S = \{s_{\lambda,\lambda'}\}$ with $s_{\lambda,\lambda'}[(A, B)] = 1$. This outcome is depicted in Figure 2.

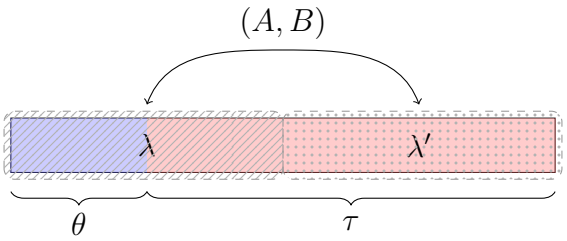

Figure 2: The matching profile in Example 5.

We verify that (Λ, p, q, μ, S) is Bayes-Nash stable in Appendix A.2.1. Intuitively, while type-θ_λ agents might attempt to target one another and propose the efficient strategy pair (A, B), agents designated to play B are reluctant to carry out the deviation. This is because the non-targeted type-τ_λ agents may also join the deviation and play B, resulting in a utility strictly lower than the status quo for the type-θ_λ agents. Notably, although (A, B) is efficient, type θ fares strictly worse than type τ in terms of average material payoffs. ♢

In Example 5, we constructed a Bayes-Nash stable outcome where the homophilic efficient type, as the minority in the population, performs worse than another type. The following proposition strengthens this observation by showing that any homophilic efficient type is dominated in evolution under incomplete information, as long as all efficient outcomes generate asymmetric material payoffs.[25]

Proposition 5. *With incomplete information, if $\pi(\widetilde{x},\widetilde{y}) \neq \pi(\widetilde{y},\widetilde{x})$ for every efficient strategy pair $(\widetilde{x},\widetilde{y})$, then any homophilic efficient type is evolutionarily unstable.*

The proof of Proposition 5 amounts to generalizing the insights from Example 5. In particular, we construct a preference type that can extricate itself from a disadvantageous position when matched with the homophilic efficient type, and discourage the latter from doing the same in a reversed situation. The first feature ensures that the constructed type receives a weakly higher average material payoff than the homophilic efficient type does across all Bayes-Nash stable outcomes, while the second feature guarantees that the inequality is sometimes strict.

In contrast, the next proposition shows that the parochial efficient type stands out even with incomplete information.

Proposition 6. *With incomplete information, the parochial efficient type is neutrally stable. Moreover, it is evolutionarily stable against any type that exhibits same-type inefficiency.*

To gain a better understanding of the stark difference between Propositions 5 and 6, it is helpful to examine the underlying logic of Example 5. In the example, the homophilic efficient agents cannot avoid unfavorable outcomes by carrying out pairwise deviations because they are concerned about the response of the type-τ agents, who may well join the deviation and behave in a way that reduces their utilities. For the parochial efficient agents, however, the behaviors of the type-τ agents *do not* matter, and matching with their own kind becomes the first priority when contemplating deviations from the status quo. This gives the parochial efficient type an incentive to break away from the disadvantageous position.

Example 6 (Example 5 revisited)**.** Consider the population state and the outcome in Example 5, with the only change that θ is now parochial efficient. The type-θ_λ agents derive a utility of 0 in the status quo. Thus, two type-θ_λ agents can form a blocking pair by targeting each other and proposing a strategy pair (A, B): For a type-θ_λ agent facing any deviation plan $(D_\lambda, \boldsymbol{x}_\lambda)$ such that $\theta \in D_\lambda$ and $\boldsymbol{x}_\lambda(\theta) = B$ (A, respectively), she plays A (B, respectively) as a best response and receives *at least* a utility of $4q_\lambda[\theta] > 0$, regardless of the behavior of

[25] Proposition 11 in Online Appendix O.2 provides another condition under which the homophilic efficient types are not neutrally stable, which further demonstrates that these preferences are not favored by evolutionary forces.

the type-τ_λ agents (i.e. whether they participate in the deviation, and if so, what strategies they play). ♢

Finally, we consider the evolutionary stability of types that exhibit selfishness. In Online Appendix O.4.3, we provide an example showing that the parochial selfish type may not be neutrally stable under incomplete information even if all strategy pairs in NE^{lb}_{π} are efficient. However, by imposing a stronger condition on the material game—where all strategy pairs in NE_π are efficient—the parochial selfish type can prevail in preference evolution.

Proposition 7. *With incomplete information, if all strategy pairs in NE_π are efficient, then the parochial selfish type is neutrally stable and evolutionarily stable against any type that exhibits same-type inefficiency.*

5 Discussions

5.1 Polymorphism

Thus far, we have focused on the stability of a monomorphic population. In this section, we extend the framework to accommodate polymorphic populations, i.e. populations consisting of multiple preference types. For simplicity, we assume complete information in this extension, although considering incomplete information is conceptually straightforward.

Let $\nu \in \Delta(\Theta)$ denote a **population distribution** with finite support in Θ and write $\Theta_\nu = \text{supp}(\nu)$. For each type $\theta \in \Theta_\nu$, the mass of type-θ agents in the population is denoted by $\nu[\theta] > 0$. The definition of an outcome (μ, S) under a population distribution ν naturally extends from the monomorphic case. Specifically, $\mu = (\mu_\theta)$ is now a vector of distributions where $\mu_\theta \in \Delta(\Theta_\nu)$ specify how agents match for each type $\theta \in \Theta_\nu$. These distributions satisfy the consistency condition:

$$\nu[\theta]\mu_\theta[\theta'] = \nu[\theta']\mu_{\theta'}[\theta], \text{ for all } \theta, \theta' \in \Theta_\nu.$$

The notions of blocking pairs and Nash stability for polymorphic populations are generalized directly from Definitions 1 and 2, with the only adjustment of replacing $\{\theta, \tau\}$ with Θ_ν. We establish existence of Nash stable outcomes in this general setting, which utilizes an existence result of stable matchings in large markets by Carmona and Laohakunakorn (2024).

Proposition 8. *Under any population distribution ν, there exists a Nash stable outcome.*

Given a Nash stable outcome (μ, S) under a population distribution ν, for each $\theta \in \Theta_\nu$,

the average material payoff for type-θ agents is given by

$$G_\theta(\mu, S) = \sum_{t \in \Theta_\nu} \mu_\theta[t] \int_{(x,y)\in\mathcal{X}^2} \pi(x,y)\, ds_{\theta,t}\,.$$

In this section, we follow Dekel et al. (2007) to examine neutral stability of a population distribution against any other preference type. Assuming that mutations are rare, with the population able to fully adjust before the next mutation occurs (Weibull, 1995), we focus on a local notion of neutral stability and impose an upper bound on the mass of the mutant type. In other words, the population distribution is always considered as the incumbent. This restriction is purely for interpretational purposes, and relaxing this upper bound does not affect our analysis mathematically.

Definition 14. A population distribution ν is **locally neutrally stable** if there exists an $\bar{\varepsilon} > 0$ such that for every $\tau \in \Theta$, $\varepsilon \in (0, \bar{\varepsilon})$, and Nash stable outcome $(\widetilde{\mu}, \widetilde{S})$ under the mixed population distribution $\widetilde{\nu} = (1-\varepsilon)\nu + \varepsilon\delta_\tau$, we have $G_\theta(\widetilde{\mu}, \widetilde{S}) \geqslant G_\tau(\widetilde{\mu}, \widetilde{S})$ for all $\theta \in \Theta_\nu$.[26]

The definition of a locally neutrally stable population distribution ν generalizes Definition 5. For any mutant type τ, all types in the support of ν must perform weakly better than τ in all Nash stable outcomes under any mixture of ν and τ, given that the proportion of type τ does not exceed a certain level. This ensures that no mutation has the ability to drive out an incumbent type in the population distribution ν. We now provide necessary conditions that describe crucial properties of population distributions that are locally neutrally stable.

Proposition 9. *Suppose the population distribution ν is locally neutrally stable.*

(i) For any Nash stable outcome (μ, S) under ν, $G_\theta(\mu, S) = G_{\theta'}(\mu, S)$ for all $\theta, \theta' \in \Theta_\nu$.

(ii) For any Nash stable outcome (μ, S) under ν such that $\mu_\theta[\theta'] > 0$ and $(x, y) \in \operatorname{supp}(s_{\theta,\theta'})$, the strategy pair (x, y) must be efficient. Moreover, either $\theta = \theta'$ or $\pi(x, y) = \pi(y, x)$.

(iii) If $\pi(\widetilde{x}, \widetilde{y}) \neq \pi(\widetilde{y}, \widetilde{x})$ for every efficient strategy pair $(\widetilde{x}, \widetilde{y})$, then for each $\theta \in \Theta_\nu$, θ does not exhibit same-type inefficiency and $u_\theta(x, y, t)$ cannot be constant in t.

Part (i) of Proposition 9 means that a locally neutrally stable ν should itself be *balanced*. If a population distribution is not balanced, some preference types will have higher fitness than others, leading natural selection to alter the distribution even before considering mutations. Dekel et al. (2007) assume balancedness when defining evolutionary stability of a polymorphic population; here, we show that it is implied by our definition. Part (ii) demonstrates that efficient play across all Nash stable outcomes is necessary for a locally neutrally stable ν. The

[26] We write $\delta_\tau \in \Delta(\Theta)$ for the Dirac measure that assigns probability one to type τ.

intuition is straightforward: Suppose the population with inefficient play is facing a parochial efficient mutant type, then in the post-entry population, some incumbent type must earn a lower average material payoff than the mutant type who separates itself from the incumbents and plays efficiently. Moreover, part (ii) indicates that any Nash stable outcome must satisfy a form of symmetry: Cross-type matches can arise only when two sides receive the same material payoff. Finally, part (iii) says that, when efficient outcomes of the material game are asymmetric, all types in a neutrally stable population distribution must play efficiently in same-type matches and exhibit plasticity. It can be viewed as an extension of Proposition 2.

While the criteria for local neutral stability may appear difficult to meet, the following result provides a sufficient condition.

Proposition 10. *If Θ_ν consists of homophilic or parochial efficient types (or both), then ν is locally neutrally stable.*

This positive result should be anticipated. If all types in the population are either homophilic or parochial efficient, then a perfectly assortative matching occurs in any post-entry population. Each type in Θ_ν matches with its own kind and derives the same average material payoff $\frac{M}{2}$. Moreover, any mutant type τ will be excluded from interacting with the types in Θ_ν and thus can receive an average material payoff at most equal to $\frac{M}{2}$.

It is natural to ask if a locally neutrally stable ν can contain types that are not homophilic or parochial efficient. For example, one may wonder whether heterophilic types, i.e. distinct types that prefer to interact with each other, can persist in evolution. Proposition 9 part (ii) suggests that this can happen only when the material game admits an efficient strategy pair that yields equal material payoffs. We illustrate this possibility in the example below.

Example 7. Consider again the prisoners' dilemma material game in Example 4, which is reproduced below:

	A	B
A	$4,4$	$0,5$
B	$5,0$	$3,3$

Let ν denote a population distribution that contains two types $\Theta_\nu = \{\theta, \theta'\}$ and $\nu[\theta] = \nu[\theta'] = \frac{1}{2}$. The utility functions of θ and θ' are given by, respectively,

$$u_\theta(x,y,t) = \begin{cases} 10 \cdot \mathbb{1}_{\{x=A\}} & \text{if } t = \theta', \\ \pi(x,y) & \text{if } t \neq \theta', \end{cases} \quad \text{and} \quad u_{\theta'}(x,y,t) = \begin{cases} 10 \cdot \mathbb{1}_{\{x=A\}} & \text{if } t = \theta, \\ \pi(x,y) & \text{if } t \neq \theta. \end{cases}$$

We now examine the local neutral stability of population distribution ν. Consider a

mutant type τ, an $\varepsilon > 0$ sufficiently small, and the post-entry population $\widetilde{\nu} = (1-\varepsilon)\nu + \varepsilon\delta_\tau$. There are two cases:

- If $\tau \notin \Theta_\nu$, then for any Nash stable outcome $(\widetilde{\mu}, \widetilde{S})$ under $\widetilde{\nu}$, we must have $\widetilde{\mu}_\theta[\theta'] = \widetilde{\mu}_{\theta'}[\theta] = 1$ and $\widetilde{s}_{\theta,\theta'}[(A,A)] = 1$. These imply that $G_\theta(\widetilde{\mu}, \widetilde{S}) = G_{\theta'}(\widetilde{\mu}, \widetilde{S}) = 4 \geqslant G_\tau(\widetilde{\mu}, \widetilde{S})$.
- If $\tau \in \Theta_\nu$, suppose $\tau = \theta$ without loss. Then any Nash stable outcome $(\widetilde{\mu}, \widetilde{S})$ under $\widetilde{\nu}$ must satisfy: (i) $\widetilde{\mu}_\theta[\theta] = \frac{2\varepsilon}{1+\varepsilon}$, $\widetilde{\mu}_\theta[\theta'] = \frac{1-\varepsilon}{1+\varepsilon}$, and $\widetilde{\mu}_{\theta'}[\theta] = 1$; (ii) $\widetilde{s}_{\theta,\theta}[(B,B)] = \widetilde{s}_{\theta,\theta'}[(A,A)] = 1$. Therefore, we have $G_\theta(\widetilde{\mu}, \widetilde{S}) < 4 = G_{\theta'}(\widetilde{\mu}, \widetilde{S})$, which is consistent with Definition 14. This means the relatively more abundant type θ will decrease in mass, and the population will revert back to ν under evolutionary forces.

Therefore, the population distribution ν consisting of heterophilic types is locally neutrally stable. It is interesting to note that both types θ and θ' exhibit same-type inefficiency. However, this mere fact does not render ν unstable, as these underlying types can secure the highest average material payoff due to heterophily. ♢

5.2 Empirical Relevancy

In experimental studies, there is limited evidence supporting the notion that people have a preference for efficiency. Charness and Rabin (2002) and Engelmann and Strobel (2004) provide some support, but it is not conclusive. A common design feature of most experimental studies is that subjects are paired or grouped exogenously and randomly, making them well-suited for examining preferences that develop in environments with random matching. In contrast, our study focuses on preferences that evolve under endogenous partner selection. However, no established experimental design exists to effectively test such preferences.

A sizable experimental literature examines how subjects' behavior in games is affected when partner choice is allowed (Ehrhart and Keser, 1999; Hauk and Nagel, 2001; Gächter and Thöni, 2005; Page et al., 2005; Gunnthorsdottir et al., 2010; Ahn et al., 2009; Slonim and Garbarino, 2008; Grimm and Mengel, 2009; Brekke et al., 2011; Rand et al., 2011; Aimone et al., 2013; Charness and Yang, 2014; Gürerk et al., 2014; Riedl et al., 2016; Guido et al., 2019, among many others) Various protocols for partner choice, including migrations across groups (possibly with different institutional arrangements or signaling values), unilateral/bilateral consent to form pairs or links to neighbors on a network, free or restricted unilateral entry/exit, voting to expel group members, voting to merge groups, matching algorithms based on elicited preferences, have been implemented in games such as prisoner's dilemmas, public good games, trust games, dictator games, and weakest-link games. Most of these studies find that partner choice is effective in promoting and sustaining cooperation or coordination on the efficient equilibrium by allowing like-minded subjects to associate with each other and

protect themselves from outsiders. In addition, in several experiments on social dilemmas (Coricelli et al., 2004; Burlando and Guala, 2005; de Oliveira et al., 2015), subjects' levels of cooperativeness are first elicited and then they are grouped/paired assortatively by the experimenters. These experiments show that exogenous sorting also substantially increases cooperation.

The findings in the literature on partner choice and exogenous grouping/pairing according to types are encouraging, as they indicate that sorting leads to higher levels of cooperation, trust, and altruism. To test if our homophilic-efficient types are empirically relevant, we can borrow elements from this literature. Eliciting subjects' preference types ex-ante and grouping/pairing them accordingly would not work for our purpose because homophilic-efficient types would not necessarily play an efficient strategy profile with strangers. Therefore, an endogenous partner choice paradigm should be used. We envision that mutual consent to form pairs with the possibility for the subjects to communicate their intended play would mimic blocking in our model. Such a partner choice protocol may be effective in sorting subjects according to types. The underlying experimental game should be some social dilemma game with a large set of strategies available and features a non-Nash efficient strategy profile. The richness of the set of strategies would give room for information revelation and the efficient strategy profile being non-Nash provides ground for efficient-preference types to flourish. Once subjects have stabilized their groupings/pairings and behavior in the game, we can then elicit their other-regarding preferences toward their own group members and other groups. By doing this, we conjecture that a higher incidence of preference for efficiency toward in-group members can be observed.

5.3 Philosophical Implications

In this section, we briefly explain the philosophical meanings of various preference types identified in this paper and compare them with those in existing literature.

First, we argue that the combination of efficient play and homophily carries significant moral significance. This assertion is substantiated by two primary reasons. First, the preference for efficiency allows individuals to prioritize mutual benefits over individual material gains. Second, the homophily manifested by our key preference types makes them all possess a fixed-point feature, which involves infinite recursive reasoning about an agent's preferences towards the opponent's preferences. For example, an agent has parochial efficient preferences if she maximizes total material payoffs and derives a positive utility only when matched with another agent, who maximizes total material payoffs and derives a positive utility only when matched with another agent, who maximizes total material payoffs and so on. Hence, a collective sentiment of "we" emerges. Our paper thus echoes the existing

literature in evolutionary psychology and anthropology by highlighting the potential influence of partner choice on the development of morality (see, for example, Baumard et al. (2013) and Tomasello (2016)).

Second, we make a comparison between the preference for efficiency and *Kantian* preference type, which has been discussed in Alger and Weibull (2013). The latter is evolutionary stable provided that the exogenous matching process's degree of assortativity is 1, i.e., positive assortative matching. The preference for efficiency, akin to most of the distributional social preferences explored in economics, is rooted in consequentialist motivations. Hypothetical imperatives, preferences over strategies due to their consequences, characterize this preference. In contrast, the *Kantian* preference is represented by the utility function $u(x) = \pi(x, x)$, implying that an agent assesses different courses of action by considering their own material payoff if the course of action were universalized to all other agents (Alger, 2022). It is characterized by categorical imperatives, preferences over strategies irrespective of their consequences, because a *Kantian* agent does not care about what other agents choose in the underlying game.[27] In games where a symmetric efficient strategy profile exists, it may not be possible to distinguish the preference for efficiency from the *Kantian* preference based on observable behavior. Nevertheless, when all the efficient strategy profiles are asymmetric, two matched agents with a preference for efficiency would obtain a greater total material payoff than two matched *Kantian* agents.

Finally, we make a comparison between the two variants of homophily in this paper. One may be inclined to view the parochial variant as the limit case of the weaker homophilic variant. However, this supposition is untenable not only mathematically (the parochial variant is lexicographic while the weak variant is not), but also philosophically. The weak variant of homophily is consequentialist, as the preference for matching with one's own type depends on the outcome of the underlying game. In contrast, the parochial variant is deontological, preferring to interact with agents of the same type regardless of the game's consequences.

6 Conclusion

In this paper, we consider preference evolution with endogenous matching by marrying the concepts of stable matching and equilibrium play. We find that the primary forces driving preference evolution are homophily and a preference for efficiency. Specifically, homophily leads to positive assortative matching, while a preference for efficiency drives the efficient play. Preferences that combine these two traits may have a fitness advantage over other preferences. Our results hold under both complete and incomplete information, although

[27] Kant (1785) refers to this as deontological motivations (Chen and Schonger, 2022).

only a strong form of homophily survives in the latter case.

There are numerous intriguing avenues for extending our work, some of which we briefly discuss here. In this paper, we take a static approach and define stability as a reduced-form outcome of an adjustment process. This, however, does not capture more intricate long-run relationships. An alternative approach is to consider a dynamic model in which match-and-play follows a history-dependent process. Ali and Liu (2025) develop a framework and solution concept to study such repeated coalitional behaviors. They show that when coalition members have perfectly aligned preferences, they collectively aim to attain the best possible outcome according to their shared objectives. For two agents with efficient preferences, their interests are indeed perfectly aligned, and the best outcome is to maximize their total material payoffs. Based on this insight, we conjecture that homophilic and parochial efficient preferences continue to prevail in a model that appropriately incorporates dynamic considerations.

Our model of incomplete information only allows for the identification of unobservable types through strategic interactions in the matching-to-interact process. However, in real-life situations, individuals may employ costly signals to reveal their types to others, which could be either strategic or genetic. Despite the issue of mimicry and deception associated with signaling, scholars have put forth the argument that certain emotions and physical states, such as uncontrollable anger or blushing, can serve as sincere indications of one's preferences (Frank, 1987, 1988; Hirshleifer, 2001). See Alger and Weibull (2019) and Alger (2022) for more discussions. Equipping individuals in the matching process with the ability to send and detect signals (Hopkins, 2014; Heller and Mohlin, 2019) may alter the matching patterns and consequently the evolution of preferences.

In order to comprehend how human preferences evolve over time, it is important to acknowledge the role of institutions. These entities can have a significant impact on people's behavior during social interactions by modifying the material benefits of the game being played and adjusting their motivations for matching through various policy instruments, such as tax and subsidies (Hiller and Touré, 2021), the protection of property rights (Bisin and Verdier, 2021), the establishment and maintenance of religious infrastructures (Bisin et al., 2021), and plans for segregation and integration (Wu, 2017). Institutions are endogenous because they are collectively determined by individuals, and as a result, preferences and institutions naturally co-evolve. A potential avenue for future research is to incorporate the approach suggested by Bisin and Verdier (2024) for modeling endogenous institutions into models of preference evolution.

A Omitted Proofs

A.1 Proofs for Section 3: Complete Information

A.1.1 Proof of Lemma 1

For the "only if" part, suppose $s'_{t,t}[NE_t^{lb}] < 1$. (If $s'_{t,t}$ is not exchangeable, S' is not a strategy profile by definition.) In other words, there exists some (x, y) such that $(x, y) \notin NE_t^{lb}$ and $(x, y) \in \text{supp}(s'_{t,t})$. Since $(x, y) \notin NE_t^{lb}$, there exists $(x^*, y^*) \in NE_t$ such that

$$\min\{u_t(x^*, y^*, t), u_t(y^*, x^*, t)\} > \min\{u_t(x, y, t), u_t(y, x, t)\}. \tag{5}$$

Therefore, two type-θ agents who are "losers" (i.e. obtain less utility) in the same-type matches can form a blocking pair and coordinate on the strategy pair (x^*, y^*). Formally, letting $u_t(x, y, t) \leqslant u_t(y, x, t)$ without loss of generality, we have

(i) $\mu_t[t] > 0$ and $(x, y) \in \text{supp}(s_{t,t})$;

(ii) $x^* \in \arg\max_{x \in \mathcal{X}} u_t(x, y^*, t)$ and $y^* \in \arg\max_{y \in \mathcal{X}} u_t(y, x^*, t)$;

(iii) $u_t(x^*, y^*, t) > u_t(x, y, t)$ and $u_t(y^*, x^*, t) > u_t(x, y, t)$.

Condition (i) comes from assumption; condition (ii) is a restatement of $(x^*, y^*) \in NE_t$; and condition (iii) is due to inequality (5). Thus, (μ, S') fails external stability, so we have a contradiction.

For the "if" part, suppose $s'_{t,t}$ is exchangeable, $s'_{t,t}[NE_t^{lb}] = 1$, and let S' be obtained by substituting $s'_{t,t}$ for $s_{t,t}$ in S, i.e. $S' = \{s'_{t,t}, s_{t,t'}, s_{t',t}, s_{t',t'}\}$ where $t' \neq t$. By contradiction, suppose (μ, S') is not Nash stable. If the blocking pair involves only type-t' agents or type-t agents in the cross-type matches, then the same blocking pair is viable in (μ, S) because the partners and strategy pairs are the same for those agents across two outcomes. If the blocking pair involves any type-t agent who has a type-t partner, then there must exist a blocking pair in (μ, S). This is because the deviating type-t agent in (μ, S') obtains a weakly higher utility than the "losers" in (μ, S) as $s'_{t,t}[NE_t^{lb}] = 1$, and these "losers" in (μ, S) have positive mass since $\mu_t[t] > 0$. Thus, the conditions for a blocking pair continue to hold. We have a contradiction in either case.

A.1.2 Proof of Proposition 1

We consider the non-trivial case where total material payoffs are not constant across all strategy pairs in the material game Γ. First observe that if θ is homophilic efficient or parochial efficient, any efficient strategy pair $(\widetilde{x}, \widetilde{y})$ constitutes a Nash equilibrium between two type-θ agents, i.e. $(\widetilde{x}, \widetilde{y}) \in NE_\theta$. This is because any unilateral deviation from an efficient

strategy pair cannot improve the total material payoff. Moreover, the set of loser-best Nash equilibria NE_θ^{lb} is exactly the set of efficient strategy pairs. To see this, simply note that the utility of a type-θ agent is equal to the total material payoff (or its monotone transformation), so any inefficient strategy pair in NE_θ leads to a strictly lower utility for both type-θ agents in a match.

Consider an arbitrary type $\tau \in \Theta$ different from θ. For $\varepsilon \in (0,1)$, take any Nash stable outcome (μ, S) in state $(\theta, \tau, \varepsilon)$. We next show that μ must be perfectly assortative, i.e. $\mu_\theta[\theta] = \mu_\tau[\tau] = 1$. By contradiction, suppose $\mu_\theta[\tau] > 0$. Then two type-θ agents in the cross-type matches can form a blocking pair and benefit from playing any efficient strategy pair $(\widetilde{x}, \widetilde{y})$ since for any $(x, y) \in \text{supp}(s_{\theta,\tau})$, we have $\pi(\widetilde{x}, \widetilde{y}) + \pi(\widetilde{y}, \widetilde{x}) + \alpha > \pi(x, y) + \pi(y, x)$ for the α-homophilic efficient type and $\pi(\widetilde{x}, \widetilde{y}) + \pi(\widetilde{y}, \widetilde{x}) > 0$ for the parochial efficient type. Since the efficient strategy pair $(\widetilde{x}, \widetilde{y})$ is indeed a Nash equilibrium between two type-θ agents, external stability is violated and (μ, S) cannot be Nash stable.

We now argue that type θ receives a weakly higher average material payoff than type τ in (μ, S). By Lemma 1, we must have $s_{\theta,\theta}[NE_\theta^{lb}] = 1$. Moreover, we have argued that NE_θ^{lb} is the set of efficient strategy pairs. Therefore,

$$\begin{aligned} G_\theta(\mu, S) &= \int_{(x,y)\in\mathcal{X}^2} \pi(x, y)\, ds_{\theta,\theta} \\ &= \frac{1}{2} \int_{(x,y)\in\mathcal{X}^2} [\pi(x, y) + \pi(y, x)]\, ds_{\theta,\theta} \\ &\geqslant \frac{1}{2} \int_{(x,y)\in\mathcal{X}^2} [\pi(x, y) + \pi(y, x)]\, ds_{\tau,\tau} \\ &= \int_{(x,y)\in\mathcal{X}^2} \pi(x, y)\, ds_{\tau,\tau} \\ &= G_\tau(\mu, S), \end{aligned}$$

where the second and the second-to-last equalities are due to the exchangeability of $s_{\theta,\theta}$ and $s_{\tau,\tau}$, and the inequality is because any strategy pair in the support of $s_{\theta,\theta}$ maximizes the total material payoff. We can conclude that the preference type θ is neutrally stable.

Now assume that type τ exhibits same-type inefficiency. By definition, write $(\widehat{x}, \widehat{y}) \in NE_\tau^{lb}$ for the inefficient strategy pair and let $\widehat{s}_{\tau,\tau}[(\widehat{x}, \widehat{y})] = \widehat{s}_{\tau,\tau}[(\widehat{y}, \widehat{x})] = \frac{1}{2}$. Then by Lemma 1, $(\mu, \widehat{S})$ is also Nash stable, where $\widehat{S}$ is obtained by substituting $\widehat{s}_{\tau,\tau}$ for $s_{\tau,\tau}$. Because $(\widehat{x}, \widehat{y})$ is inefficient, the inequality is above is strict, i.e. $G_\theta(\mu, \widehat{S}) > G_\tau(\mu, \widehat{S})$. Therefore, θ is evolutionarily stable against τ.

A.1.3 Proof of Proposition 2

(i) By Proposition 1, if θ exhibits same-type inefficiency, the parochial efficient type is evolutionarily stable against θ. Thus, θ is evolutionarily unstable by definition.

(ii) Let θ be a type such that $u_\theta(x,y,t) = f(x,y)$. If θ exhibits same-type inefficiency, then part (i) applies. Now suppose all strategy pairs in NE_θ^{lb} are efficient. Consider a type τ that has the following utility function (defined on $X^2 \times \Theta$ and extended to $\mathcal{X}^2 \times \Theta$):

$$u_\tau(x,y,t) = \begin{cases} \pi(x,y) + \pi(y,x) & \text{if } t = \tau, \\ [\pi(x,y) + \pi(y,x)] \cdot \mathbb{1}_{\{\pi(x,y) \geqslant \pi(y,x)\}} & \text{if } t \neq \tau. \end{cases}$$

When matched with her own kind, a type-τ agent cares about efficiency. When matched with a type-θ agent, however, she derives utility only if she can earn a higher material payoff than her partner. For $\varepsilon \in (0,1)$, take any Nash stable outcome (μ, S) at state $(\theta, \tau, \varepsilon)$. First, note that the set NE_τ^{lb} is exactly the set of efficient strategy pairs because the utility of type-τ agents when matched with each other equals the total material payoff. Therefore, if $\mu_\tau[\tau] > 0$, $s_{\tau,\tau}$ attaches probability one to efficient strategy pairs by Lemma 1; the same holds for $s_{\theta,\theta}$ if $\mu_\theta[\theta] > 0$. If $\mu_\tau[\theta] > 0$ and $(x,y) \in \operatorname{supp}(s_{\tau,\theta})$, then it must be that $(x,y) \in \mathcal{X}^2$ is efficient and $\pi(x,y) \geqslant \pi(y,x)$. For if not, two type-$\tau$ agents in cross-type matches can form a blocking pair and benefit from playing any efficient strategy pair due to the form of their utility function. Let M denote the maximum total material payoff, and thus $\pi(x,y) \geqslant \frac{1}{2}M \geqslant \pi(y,x)$ for all $(x,y) \in \operatorname{supp}(s_{\tau,\theta})$. Therefore, by exchangeability, we have

$$\int_{(x,y)\in\mathcal{X}^2} \pi(x,y)\, ds_{\tau,\theta} \geqslant \frac{1}{2}M \geqslant \int_{(x,y)\in\mathcal{X}^2} \pi(y,x)\, ds_{\tau,\theta} = \int_{(x,y)\in\mathcal{X}^2} \pi(x,y)\, ds_{\theta,\tau},$$

which further implies

$$\begin{aligned} G_\tau(\mu, S) &= \mu_\tau[\tau] \int_{(x,y)\in\mathcal{X}^2} \pi(x,y)\, ds_{\tau,\tau} + \mu_\tau[\theta] \int_{(x,y)\in\mathcal{X}^2} \pi(x,y)\, ds_{\tau,\theta} \\ &\geqslant \frac{1}{2}M \\ &\geqslant \mu_\theta[\theta] \int_{(x,y)\in\mathcal{X}^2} \pi(x,y)\, ds_{\theta,\theta} + \mu_\theta[\tau] \int_{(x,y)\in\mathcal{X}^2} \pi(x,y)\, ds_{\theta,\tau} \\ &= G_\theta(\mu, S). \end{aligned}$$

We now argue that the inequality is strict for some Nash stable outcome $(\mu, \widetilde{S})$. This is true when $\mu_\tau[\theta] > 0$, $\widetilde{s}_{\theta,\theta}[NE_\theta^{lb}] = \widetilde{s}_{\tau,\tau}[NE_\tau^{lb}] = 1$, and $\widetilde{s}_{\tau,\theta}[(\widetilde{x},\widetilde{y})] = 1$, where $(\widetilde{x},\widetilde{y}) \in NE_\theta^{lb}$ is efficient and $\pi(\widetilde{x},\widetilde{y}) > \pi(\widetilde{y},\widetilde{x})$. The existence of such an outcome is guaranteed by the

assumption that all strategy pairs in NE_θ^{lb} are efficient, and all efficient strategy pairs result in unbalanced material payoffs. Internal stability is satisfied because type-τ agents do not care about their partner's type and $(\widetilde{x}, \widetilde{y}) \in NE_\theta$. To check external stability, first note that all type-τ agents already obtain their highest possible utility, so they do not participate in any blocking pair. If two type-θ agents form a blocking pair and coordinate on some Nash equilibrium $(\widehat{x}, \widehat{y}) \in NE_\theta^{lb}$, we must have

$$\min\{\pi(\widehat{x}, \widehat{y}), \pi(\widehat{y}, \widehat{x})\} > \pi(\widetilde{y}, \widetilde{x}) = \min\{\pi(\widetilde{x}, \widetilde{y}), \pi(\widetilde{y}, \widetilde{x})\},$$

contradicting the assumption that $(\widetilde{x}, \widetilde{y}) \in NE_\theta^{lb}$. Therefore, $(\mu, \widetilde{S})$ is a Nash stable outcome in which we have $G_\theta(\mu, \widehat{S}) > \frac{1}{2}M > G_\tau(\mu, \widehat{S})$. Hence, type θ is evolutionarily unstable.

A.1.4 Proof of Proposition 3

We consider the non-trivial case where total material payoffs are not constant across all strategy pairs in the material game Γ. For part (i), write θ for the α-homophilic selfish type and take $\alpha > \max_{(x,y)\in\mathcal{X}^2} \pi(x, y)$. Therefore, if some type-θ agents are matched with type-τ ones (i.e. $\mu_\theta[\tau] > 0$), they can always form a blocking pair and play any Nash equilibrium between themselves. This means any Nash stable outcome (μ, S) should be perfectly assortative, $\mu_\theta[\theta] = \mu_\tau[\tau] = 1$. Neutral stability then follows from noting that all strategy pairs in $NE_\theta^{lb} = NE_\pi^{lb}$ are efficient by assumption and applying Lemma 1 to type θ. In addition, evolutionary stability when τ exhibits same-type inefficiency follows from applying Lemma 1 again to τ.

For part (ii), write θ for the parochial selfish type; therefore, $NE_\theta^{lb} = NE_\pi^{lb}$. A type-$\theta$ agent derives zero utility when matched with a type-τ agent. Thus, if $\mu_\theta[\tau] > 0$, type-θ agents in cross-type matches can always form a blocking pair with each other and coordinate on any Nash equilibrium strategy pair $(x, y) \in NE_\pi^{lb}$ which ensures positive utilities for both agents. To see this, first note that $\pi(x, y) \geqslant 0$ and $\pi(y, x) \geqslant 0$ by assumption. At least one of the inequalities is strict because (x, y) is efficient. If both are strict, we are done; if only one is strict, then there exists a symmetric mixed strategy Nash equilibrium where both agents obtain strictly positive utility, contradicting the fact that $(x, y) \in NE_\pi^{lb}$. Therefore, we have $\mu_\theta[\theta] = \mu_\tau[\tau] = 1$ in any Nash stable outcome (μ, S). As in part (i), neutral stability follows from noting that all strategy pairs in NE_π^{lb} are efficient and applying Lemma 1 to θ. In addition, evolutionary stability when τ exhibits same-type inefficiency follows from applying Lemma 1 again to τ.

A.1.5 Proof of Proposition 4

First consider the parochial selfish type. We argue that all strategy pairs in NE^{lb}_{π} must be efficient. To see this, take any $(x, y) \in NE^{lb}_{\pi}$, we must have

$$\pi(\widetilde{x}, \widetilde{x}) + \pi(\widetilde{x}, \widetilde{x}) \geqslant \pi(x, y) + \pi(y, x) \geqslant \pi(\widetilde{x}, \widetilde{x}) + \pi(\widetilde{x}, \widetilde{x}).$$

The first inequality is because $(\widetilde{x}, \widetilde{x})$ is efficient. For the second inequality, suppose instead $\pi(x, y) + \pi(y, x) < \pi(\widetilde{x}, \widetilde{x}) + \pi(\widetilde{x}, \widetilde{x})$. This in turn means $\min\{\pi(x, y), \pi(y, x)\} < \pi(\widetilde{x}, \widetilde{x})$. Because $(\widetilde{x}, \widetilde{x})$ is a Nash equilibrium by assumption, the inequality implies that (x, y) cannot be a loser-best Nash equilibrium, a contradiction. We can then invoke Proposition 3 and conclude that the parochial selfish type is neutrally stable and evolutionarily stable against any type that exhibits same-type inefficiency.

For the α-homophilic selfish type with any $\alpha > 0$, we first consider the case that $\mu_\theta[\tau] = 0$. Because all strategy pairs in NE^{lb}_{θ} are efficient, applying Lemma 1 to θ ensures that $G_\theta(\mu, S) \geqslant G_\tau(\mu, S)$ for all Nash stable outcomes (μ, S). If $\mu_\theta[\tau] > 0$ in a Nash stable outcome (μ, S), take any $(x, y) \in \text{supp}(s_{\theta,\tau})$, we must have

$$\pi(x, y) > \pi(\widetilde{x}, \widetilde{x}) > \pi(y, x).$$

To see why, observe that if the first inequality does not hold, two α-homophilic selfish agents can form a blocking pair and coordinate on the Nash equilibrium $(\widetilde{x}, \widetilde{x})$; if the second inequality does not hold, we have $\pi(x, y) + \pi(y, x) > \pi(\widetilde{x}, \widetilde{x}) + \pi(\widetilde{x}, \widetilde{x})$ which contradicts the assumption that $(\widetilde{x}, \widetilde{x})$ is efficient. Therefore, since

$$\int_{(x,y)\in\mathcal{X}^2} \pi(x, y)\, ds_{\theta,\tau} > \pi(\widetilde{x}, \widetilde{x}) > \int_{(x,y)\in\mathcal{X}^2} \pi(y, x)\, ds_{\theta,\tau} = \int_{(x,y)\in\mathcal{X}^2} \pi(x, y)\, ds_{\tau,\theta},$$

$(\widetilde{x}, \widetilde{x})$ is efficient, and all strategy pairs in NE^{lb}_{θ} are efficient, we have $G_\theta(\mu, S) > G_\tau(\mu, S)$. Thus, θ is neutrally stable.

When τ exhibits same-type inefficiency, applying Lemma 1 again to τ guarantees the existence of a Nash stable outcome $(\mu, \widehat{S})$ such that $G_\theta(\mu, \widehat{S}) > G_\tau(\mu, \widehat{S})$ even in the case that $\mu_\theta[\tau] = 0$. Hence, θ is evolutionarily stable against τ.

A.2 Proofs for Section 4: Incomplete Information

In this section, we first elaborate on Example 5. Next, we establish the positive results under incomplete information, Propositions 6 and 7. Finally, we prove the negative result on homophilic efficient preferences, Proposition 5. This order is chosen because the proofs of

the positive results deliver more important economic insights, and the underlying blocking mechanism also plays a role in proving Proposition 5.

A.2.1 More on Example 5

We formally verify that (Λ, p, q, μ, S) is a Bayes-Nash equilibrium profile in Example 5.

- All type-θ_λ agents already obtain the highest possible utility when interacting with a type-τ partner. Therefore, they cannot improve their utility by targeting type-τ_λ or type-$\tau_{\lambda'}$ agents in an incomplete information blocking pair.

- For type-τ agents with either label, the only possibility of deviation is to target another type-τ agent and coordinate on a Nash equilibrium of the material game (since they are selfish when interacting with their own kind). Let us consider the pure strategy Nash equilibrium (A, B) or (B, A). For the side that is positioned to play A, the utility in the deviation is 1 if only targeted agents participate, which is no more than her current utility. This means the proposed deviation does not increase the utility of one side of the type-τ agents. The mixed strategy equilibrium can be ruled out in a similar way.

- Next, we check the case where two type-θ_λ agents target each other and propose the efficient outcome (B, A) or (A, B). Consider a type-θ_λ agent who is positioned to play B facing a deviation plan $(D, \boldsymbol{x})$ such that $D = \{\theta, \tau\}$, $\boldsymbol{x}(\theta) = A$, and $\boldsymbol{x}(\tau) = B$.[28] In this case, the type-θ_λ agent obtains $(4 + \alpha)q_\lambda[\theta]$ by playing B, which is no more than her utility 4 in the status quo since $q_\lambda[\theta] \leqslant \frac{4}{4+\alpha}$ by construction.

- Finally, suppose two type-θ_λ agents target each other and propose the inefficient equilibrium $(\frac{1}{2}A + \frac{1}{2}B, \frac{1}{2}A + \frac{1}{2}B)$. Consider a deviation plan $(D, \boldsymbol{x})$ such that $D = \{\theta, \tau\}$, $\boldsymbol{x}(\theta) = \frac{1}{2}A + \frac{1}{2}B$, and $\boldsymbol{x}(\tau) = B$. Any side of type-θ_λ agents facing this deviation plan will have a strict best response A, violating optimality of the proposed strategy $\frac{1}{2}A + \frac{1}{2}B$.

Therefore, no viable incomplete information blocking pair exists.

A.2.2 Proof of Proposition 6

We consider the non-trivial case where total material payoffs are not constant across all strategy pairs in the material game Γ. Write θ for the parochial efficient type. Consider an arbitrary preference type $\tau \in \Theta$. Take any Bayes-Nash stable outcome (Λ, p, q, μ, S). We establish the result by a sequence of lemmas.

[28] Note that B is indeed a rational and profitable play of a type-τ_λ agent who faces a type-θ_λ partner.

Lemma 2. *If $\lambda \in \Lambda$ and $q_\lambda[\tau] = 1$, then $\mu_\lambda[\lambda] = 1$.*

Proof. Suppose $\mu_\lambda[\lambda'] > 0$ for some $\lambda' \neq \lambda$. There always exists an incomplete information blocking pair between two type-$\theta_{\lambda'}$ agents who target each other and propose to play an efficient strategy pair denoted by $(\widetilde{x}, \widetilde{y})$. To see this, suppose $(x, y) \in \mathrm{supp}(s_{\lambda,\lambda'})$ and let us verify the incentives of a type-$\theta_{\lambda'}$ agent who agrees to play $\widetilde{x}$. Take any deviation plan $(D', \boldsymbol{y})$ for a label-λ' agent such that $\theta \in D'$ and $\boldsymbol{y}(\theta) = \widetilde{y}$. If $\tau \notin D'$, we have $\mathbb{E}_{q_{\lambda'}}[u_\theta(x, \boldsymbol{y}(\cdot), \cdot) \,|\, D'] = u_\theta(x, \widetilde{y}, \theta)$,

$$\begin{aligned} \widetilde{x} \in \arg\max_{x \in \mathcal{X}} u_\theta(x, \widetilde{y}, \theta), \text{ and} \\ u_\theta(\widetilde{x}, \widetilde{y}, \theta) = \pi(\widetilde{x}, \widetilde{y}) + \pi(\widetilde{y}, \widetilde{x}) > 0 = u_\theta(x, y, \lambda), \end{aligned}$$

because $(\widetilde{x}, \widetilde{y})$ is an efficient strategy pair. If $\tau \in D'$, we have

$$\begin{aligned} \widetilde{x} \in \arg\max_{x \in \mathcal{X}} u_\theta(x, \widetilde{y}, \theta) = \arg\max_{x \in \mathcal{X}} \mathbb{E}_{q_{\lambda'}}[u_\theta(x, \boldsymbol{y}(\cdot), \cdot) \,|\, D'], \text{ and} \\ \mathbb{E}_{q_{\lambda'}}[u_\theta(x, \boldsymbol{y}(\cdot), \cdot) \,|\, D'] &= q_{\lambda'}[\theta] u_\theta(\widetilde{x}, \widetilde{y}, \theta) + q_{\lambda'}[\tau] u_\theta(\widetilde{x}, \boldsymbol{y}(\tau), \tau) \\ &= q_{\lambda'}[\theta] \cdot [\pi(\widetilde{x}, \widetilde{y}) + \pi(\widetilde{y}, \widetilde{x})] \\ &> 0 = u_\theta(x, y, \lambda) \end{aligned}$$

because $u_\theta(\cdot, \cdot, \tau) = 0$ and $q_{\lambda'}[\theta] > 0$. Therefore, the type-$\theta_{\lambda'}$ agent in question is willing to participate in the deviation and play $\widetilde{x}$ as a best response. The incentives of the other side who agrees to participate and play $\widetilde{y}$ can be verified similarly. Hence, there exists an incomplete information blocking pair which contradicts the fact that (Λ, p, q, μ, S) is Bayes-Nash stable. □

Lemma 3. *If $\lambda \in \Lambda$ and $q_\lambda[\theta] > 0$, then $\mu_\lambda[\lambda] = 1$ and any strategy pair $(x, y) \in \mathrm{supp}(s_{\lambda,\lambda})$ is efficient.*

Proof. By contradiction, suppose $(x, y) \in \mathrm{supp}(s_{\lambda,\lambda})$ is inefficient or $\mu_\lambda[\lambda'] > 0$ for some $\lambda' \neq \lambda$. In the latter case, assume $q_\lambda[\theta] > q_{\lambda'}[\theta]$ without loss of generality. Then there exists an incomplete information blocking pair formed by two type-θ_λ agents who agree to play an efficient strategy pair denoted by $(\widetilde{x}, \widetilde{y})$. Formally, consider a type-θ_λ agent who agrees to play $\widetilde{x}$ and take any deviation plan $(D', \boldsymbol{y})$ for a label-λ agent such that $\theta \in D'$ and $\boldsymbol{y}(\theta) = \widetilde{y}$. If $\tau \notin D'$, we have $\mathbb{E}_{q_\lambda}[u_\theta(x, \boldsymbol{y}(\cdot), \cdot) \,|\, D'] = u_\theta(x, \widetilde{y}, \theta)$,

$$\begin{aligned} \widetilde{x} \in \arg\max_{x \in \mathcal{X}} u_\theta(x, \widetilde{y}, \theta), \text{ and} \\ u_\theta(\widetilde{x}, \widetilde{y}, \theta) = \pi(\widetilde{x}, \widetilde{y}) + \pi(\widetilde{y}, \widetilde{x}) > q_{\lambda'}[\theta] \cdot [\pi(x, y) + \pi(y, x)] = u_\theta(x, y, \lambda). \end{aligned}$$

The inequality is strict because either (x, y) is inefficient or $q_{\lambda'}[\theta] < 1$. If $\tau \in D'$, we have

$$\begin{aligned}
\widetilde{x} \in \arg\max_{x \in \mathcal{X}} u_\theta(x, \widetilde{y}, \theta) &= \arg\max_{x \in \mathcal{X}} \mathbb{E}_{q_\lambda}[u_\theta(x, \boldsymbol{y}(\cdot), \cdot) \,|\, D'], \text{ and} \\
\mathbb{E}_{q_\lambda}[u_\theta(x, \boldsymbol{y}(\cdot), \cdot) \,|\, D'] &= q_\lambda[\theta] u_\theta(\widetilde{x}, \widetilde{y}, \theta) + q_\lambda[\tau] u_\theta(\widetilde{x}, \boldsymbol{y}(\tau), \tau) \\
&= q_\lambda[\theta] \cdot [\pi(\widetilde{x}, \widetilde{y}) + \pi(\widetilde{y}, \widetilde{x})] \\
&> q_{\lambda'}[\theta] \cdot [\pi(x, y) + \pi(y, x)] \\
&= u_\theta(x, y, \lambda).
\end{aligned}$$

The inequality is strict because either (x, y) is inefficient or $q_\lambda[\theta] > q_{\lambda'}[\theta]$. The incentives of the other side who agrees to participate and play $\widetilde{y}$ can be verified similarly. Hence, there exists an incomplete information blocking pair which leads to a contradiction. □

Letting λ_τ denote the label that fully reveals type τ, i.e. $q_{\lambda_\tau}[\tau] = 1$, the lemmas above imply that

$$G_\theta(\Lambda, p, q, \mu, S) = \sum_{\lambda \in \Lambda} \frac{p_\lambda q_\lambda[\theta]}{1 - \varepsilon} \cdot \mu_\lambda[\lambda] \cdot \frac{M}{2} = \frac{M}{2} = G_\tau(\Lambda, p, q, \mu, S) \quad \text{if } \lambda_\tau \notin \Lambda,$$

and

$$\begin{aligned}
G_\theta(\Lambda, p, q, \mu, S) &\geqslant \left(1 - \frac{p_{\lambda_\tau}}{\varepsilon}\right) \cdot \frac{M}{2} + \frac{p_{\lambda_\tau}}{\varepsilon} \cdot \mu_{\lambda_\tau}[\lambda_\tau] \int_{(x,y) \in \mathcal{X}^2} \pi(x, y)\, ds_{\lambda_\tau, \lambda_\tau} \\
&= G_\tau(\Lambda, p, q, \mu, S) \quad \text{if } \lambda_\tau \in \Lambda.
\end{aligned}$$

We can conclude that type θ is neutrally stable under incomplete information.

Now suppose that type τ exhibits same-type inefficiency. Note that there always exists a Bayes-Nash stable outcome (Λ, p, q, μ, S) with $\lambda_\tau \in \Lambda$ and $p_{\lambda_\tau} > 0$ as a complete information Nash stable outcome always exists and is a special case. Whenever $p_{\lambda_\tau} > 0$, in the spirit of Lemma 1, we can construct another Bayes-Nash stable outcome $(\Lambda, p, q, \mu, \widehat{S})$ such that (x, y) is inefficient for all $(x, y) \in \operatorname{supp}(\widehat{s}_{\lambda_\tau, \lambda_\tau})$ where $\widehat{s}_{\lambda_\tau, \lambda_\tau} \in \widehat{S}$. This means the inequality above must be strict, i.e. $G_\theta(\Lambda, p, q, \mu, \widehat{S}) > G_\tau(\Lambda, p, q, \mu, \widehat{S})$. Therefore, the parochial efficient type θ is evolutionarily stable against τ.

A.2.3 Proof of Proposition 7

Most of the arguments below are similar to those in the proof of Proposition 6, so we omit some details. Write θ for the parochial selfish type and and take any Bayes-Nash stable outcome (Λ, p, q, μ, S). One can show that if $\lambda \in \Lambda$ and $q_\lambda[\tau] = 1$, then $\mu_\lambda[\lambda] = 1$. For if not,

i.e. $\mu_\lambda[\lambda'] > 0$ for some $\lambda' \neq \lambda$, there exists an incomplete information blocking pair formed by two type-$\theta_{\lambda'}$ agents who target each other and propose to play any strategy pair in NE_θ.

For agents of other labels, perfect assortativity may fail. However, whenever two different labels are matched in a Bayes-Nash stable outcome, the label that contains more type-θ agents must receive a strictly higher material payoff, as shown below.

Lemma 4. *For $\lambda \in \Lambda$, if $q_\lambda[\theta] > 0$ and $\mu_\lambda[\lambda'] > 0$, then for any strategy pair $(x, y) \in \text{supp}(s_{\lambda,\lambda'})$, we have (i) (x, y) is efficient and (ii) $\pi(x, y) > \pi(y, x)$ if $q_\lambda[\theta] > q_{\lambda'}[\theta]$.*

Proof. If $q_\lambda[\theta] > 0$ and $\mu_\lambda[\lambda'] > 0$, by the previous argument, we must have $q_{\lambda'}[\theta] > 0$. Then internal stability implies that $(x, y) \in NE_\theta = NE_\pi$ due to the form of type θ's utility function. By assumption, (x, y) is efficient.

By contradiction, assume that $q_\lambda[\theta] > q_{\lambda'}[\theta]$ and $\pi(x, y) \leqslant \pi(y, x)$. Then there exists an incomplete information blocking pair formed by two type-θ_λ agents who target each other and propose the strategy pair $(x, y) \in NE_\theta$. This is because both sides of the type-θ_λ agents can secure a payoff of at least

$$q_\lambda[\theta] \cdot \min\{\pi(x, y), \pi(y, x)\} = q_\lambda[\theta] \cdot \pi(x, y) > q_{\lambda'}[\theta] \cdot \pi(x, y)$$

conditional on their targeted partners' participation. □

Suppose there are two labels $\lambda, \lambda' \in \Lambda$ such that $\lambda \neq \lambda'$ and $\mu_\lambda[\lambda'] > 0$. Without loss of generality, we assume $q_\lambda[\theta] > q_{\lambda'}[\theta]$. By Lemma 4 and the fact that $p_\lambda \mu_\lambda[\lambda'] = p_{\lambda'} \mu_{\lambda'}[\lambda]$, the average material payoff of type θ across λ-λ' matches can be computed as

$$\begin{aligned}
&\frac{1}{q_\lambda[\theta] + q_{\lambda'}[\theta]} \left\{ q_\lambda[\theta] \int_{\mathcal{X}^2} \pi(x, y)\, ds_{\lambda,\lambda'} + q_{\lambda'}[\theta] \int_{\mathcal{X}^2} \pi(x, y)\, ds_{\lambda',\lambda} \right\} \\
= &\frac{1}{q_\lambda[\theta] + q_{\lambda'}[\theta]} \left\{ \left(\frac{q_\lambda[\theta] + q_{\lambda'}[\theta]}{2} + \frac{q_\lambda[\theta] - q_{\lambda'}[\theta]}{2} \right) \int_{\mathcal{X}^2} \pi(x, y)\, ds_{\lambda,\lambda'} + q_{\lambda'}[\theta] \int_{\mathcal{X}^2} \pi(x, y)\, ds_{\lambda',\lambda} \right\} \\
> &\frac{1}{q_\lambda[\theta] + q_{\lambda'}[\theta]} \left\{ \frac{q_\lambda[\theta] + q_{\lambda'}[\theta]}{2} \int_{\mathcal{X}^2} \pi(x, y)\, ds_{\lambda,\lambda'} + \frac{q_\lambda[\theta] - q_{\lambda'}[\theta]}{2} \int_{\mathcal{X}^2} \pi(y, x)\, ds_{\lambda,\lambda'} + q_{\lambda'}[\theta] \int_{\mathcal{X}^2} \pi(x, y)\, ds_{\lambda',\lambda} \right\} \\
= &\frac{1}{q_\lambda[\theta] + q_{\lambda'}[\theta]} \left\{ \frac{q_\lambda[\theta] + q_{\lambda'}[\theta]}{2} \int_{\mathcal{X}^2} \pi(x, y)\, ds_{\lambda,\lambda'} + \frac{q_\lambda[\theta] + q_{\lambda'}[\theta]}{2} \int_{\mathcal{X}^2} \pi(y, x)\, ds_{\lambda,\lambda'} \right\} \\
= &\frac{1}{2} \int_{\mathcal{X}^2} [\pi(x, y) + \pi(y, x)]\, ds_{\lambda,\lambda'} \\
= &\frac{1}{2} M.
\end{aligned}$$

Therefore, the average material payoff of type θ in the population must satisfy $G_\theta(\Lambda, p, q, \mu, S) \geqslant \frac{M}{2}$, where the inequality is strict if there is a positive mass of cross-label matches. By a similar argument for type τ, we have $G_\tau(\Lambda, p, q, \mu, S) \leqslant \frac{M}{2}$. This implies that type θ is neutrally stable under incomplete information.

When τ exhibits same-type inefficiency, we can follow the argument in the proof of Proposition 6 to show that θ is evolutionarily stable against τ.

A.2.4 Proof of Proposition 5

Fixing $\alpha > 0$, denote by θ the α-homophilic efficient type. Write $E \subseteq \mathcal{X}^2$ for the set of efficient strategy pairs and define two subsets of X as

$$X^+ = \{x \in X : \pi(x,y) > \pi(y,x) \text{ for some } (x,y) \in E\}, \text{ and}$$
$$X^- = \{x \in X : \pi(x,y) < \pi(y,x) \text{ for some } (x,y) \in E\}.$$

By assumption, we have $X^+ \cap X^- = \varnothing$; for if not, there will be a mixed strategy pair that is efficient and generates equal material payoffs. Pick $x^+ \in X^+$. Consider a preference type τ that has the following utility function

$$u_\tau(x,y,t) = \begin{cases} \mathbb{1}_{\{(x,y)\in E\}} & \text{if } t = \tau, \\ 2 \cdot \mathbb{1}_{\{x = x^+\}} & \text{if } t \neq \tau. \end{cases}$$

Intuitively, type-τ agents care about efficiency when playing against themselves, but prefer to play x^+ when matched with other types. We now show that τ is evolutionarily stable against θ by establishing two lemmas.

Lemma 5. *$G_\tau(\Lambda, p, q, \mu, S) \geqslant G_\theta(\Lambda, p, q, \mu, S)$ for all Bayes-Nash stable outcomes (Λ, p, q, μ, S).*

Proof. Fix an arbitrary Bayes-Nash stable outcome (Λ, p, q, μ, S). We prove this lemma by establishing the following claim: For $\lambda \in \Lambda$, if $q_\lambda[\tau] > 0$ and $\mu_\lambda[\lambda'] > 0$, then we have (i) $s_{\lambda,\lambda'}[E] = 1$ and (ii) $s_{\lambda,\lambda'}[\{(x,y) : x = x^+\}] = 1$ if $q_\lambda[\tau] > q_{\lambda'}[\tau]$.

For part (i), suppose $\mu_\lambda[\lambda'] > 0$ and there is some inefficient strategy pair $(x,y) \in \operatorname{supp}(s_{\lambda,\lambda'})$. Also suppose $q_\lambda[\tau] \geqslant q_{\lambda'}[\tau]$ without loss. In this case, we cannot have $x = x^+$; for if so, any best response y for type-$\theta_{\lambda'}$ agents must satisfy $(x,y) \in E$, contradicting the assumption that (x,y) is inefficient. Moreover, for every $\widehat{x} \in \operatorname{supp}(x)$, we have $(\widehat{x}, y) \notin E$; otherwise, $\widehat{x}$ delivers higher utility to type-θ_λ agents than x, a contradiction. The utility of these type-τ_λ agents then satisfies

$$2q_{\lambda'}[\theta] \leqslant u_\tau(x,y,\lambda') < q_{\lambda'}[\tau],$$

where the first inequality comes from the fact that a type-τ_λ agent can secure at least $2q_{\lambda'}[\theta]$ by playing x^+, and the second inequality is because x attaches positive probability to some

pure strategy other than x^+.[29] Thus, there exists an incomplete information blocking pair formed by two such type-τ_λ agents who propose to play $(x^+, y^-) \in E$. For the side that agrees to play y^-, strategy y^- is always a best response because, even if type-θ_λ agents join the deviation, we have $q_\lambda[\tau] \geqslant q_{\lambda'}[\tau] > 2q_{\lambda'}[\theta] \geqslant 2q_\lambda[\theta]$;[30] moreover, the utility obtained from the deviation is at least $q_\lambda[\tau] \geqslant q_{\lambda'}[\tau] > u_\tau(x, y, \lambda')$. On the other hand, the side that agrees to play x^+ will clearly participate, because when type-θ_λ agents join the deviation, playing x^+ is still a best response and yields an even higher utility.

For part (ii), by contradiction, suppose $q_\lambda[\tau] > q_{\lambda'}[\tau]$ and $(x, y) \in \text{supp}(s_{\lambda,\lambda'})$ where $(x, y) \in E$ and $x \neq x^+$. By internal stability, some pure strategy other than x^+ is a best response for type-τ_λ agents, so

$$2q_{\lambda'}[\theta] \leqslant u_\tau(x, y, \lambda') = q_{\lambda'}[\tau].$$

But then the two type-τ_λ agents can target each other and form an incomplete information blocking pair by proposing $(x^+, y^-) \in E$. For the side that agrees to play y^-, strategy y^- is always a best response because, even if type-θ_λ agents join the deviation, we have $q_\lambda[\tau] > q_{\lambda'}[\tau] \geqslant 2q_{\lambda'}[\theta] > 2q_\lambda[\theta]$; moreover, the utility received from the deviation is at least $q_\lambda[\tau] > q_{\lambda'}[\tau] = u_\tau(x, y, \lambda')$. On the other hand, the side that agrees to play x^+ will honor the promise for the same reason as in part (i).

One can then follow the argument in the proof of Proposition 7 to show that $G_\tau(\Lambda, p, q, \mu, S) \geqslant \frac{M}{2} \geqslant G_\theta(\Lambda, p, q, \mu, S)$, where the inequalities become strict if there is a positive mass of cross-label matches. In the next lemma, we shall show that the case of cross-label matches is indeed possible. □

Lemma 6. *$G_\tau(\Lambda, p, q, \mu, S) > G_\theta(\Lambda, p, q, \mu, S)$ for some Bayes-Nash stable outcome (Λ, p, q, μ, S).*

Proof. Let $M^p < M$ denote the highest total material payoff delivered by an inefficient pure strategy pair, i.e.

$$M^p = \max_{(x,y)\in X^2\setminus E} \pi(x, y) + \pi(y, x).$$

Moreover, let M^n be the highest total material payoff derived from an inefficient Nash equilibrium between two type-θ agents (which always exists), i.e.

$$M^n = \max_{(x,y)\in NE_\theta\setminus E} \pi(x, y) + \pi(y, x).$$

[29] Let $\widehat{x} \neq x^+$ denote this pure strategy. Then $u_\tau(x, y, \lambda') = u_\tau(\widehat{x}, y, \lambda') = q_{\lambda'}[\tau]u_\tau(\widehat{x}, y, \tau) < q_{\lambda'}[\tau]$ because we have argued that $(\widehat{x}, y) \notin E$.

[30] If type-θ_λ agents join the deviation, the type-τ_λ agent on this side obtains $q_\lambda[\tau]$ by playing y^-, while the best alternative is to play x^+ and obtain $2q_\lambda[\theta]$ because $(x^+, x^+) \notin E$.

Since NE_θ is a finite union of maximal Nash subsets (Jansen, 1981) and the total material payoff is constant on each subset due to θ's utility function, $M^n < M$ is well-defined.

Fix a population state $(\theta, \tau, \varepsilon)$ and consider an outcome (Λ, p, q, μ, S) as follows. There are three labels $\Lambda = \{\lambda_\theta, \lambda, \lambda_\tau\}$, and the latter two have equal masses $p_\lambda = p_{\lambda_\tau}$. Labels with a subscript perfectly reveal underlying types, i.e. $q_{\lambda_\theta}[\theta] = q_{\lambda_\tau}[\tau] = 1$. The proportion of type-θ agents among those with label λ satisfies

$$q_\lambda[\theta] \leqslant \min\left\{\frac{M - M^n}{\alpha}, \frac{M - M^p}{M - M^p + \alpha}\right\}.$$

Assume all label-λ agents are matched with label-λ_τ, that is, $\mu_\lambda[\lambda_\tau] = \mu_{\lambda_\tau}[\lambda] = 1$. Finally, let $(x^+, y^-) \in E$ and the strategy profile S is such that $s_{\lambda_\theta,\lambda_\theta}[E] = 1$ and $s_{\lambda,\lambda_\tau}[(y^-, x^+)] = s_{\lambda_\tau,\lambda}[(x^+, y^-)] = 1$. Note that a matching profile (Λ, p, q, μ) satisfying the conditions above is always feasible for any $\varepsilon > 0$. Figure 3 below illustrates such an outcome.

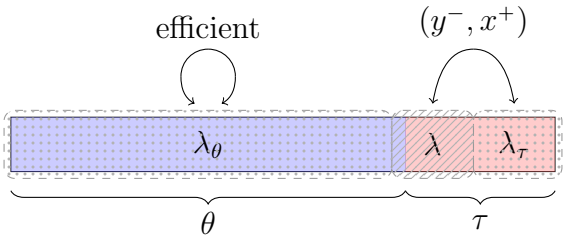

Figure 3: The matching profile for the proof of part (ii).

We argue that this outcome is Bayes-Nash stable. First, type-θ_{λ_θ} agents have no incentive to participate in a blocking pair, as they already obtain their maximum utility. Because the play between labels λ and λ_τ is already efficient, no type-θ_λ agent has an incentive to target a type-τ agent (with label λ or λ_τ) in a blocking pair. Moreover, no type-τ agent (with label λ or λ_τ) has an incentive to target another type-τ agent since her utility is no less than 1 in the status quo. Therefore, it is only left to consider blocking pairs consisting of two type-θ_λ agents who target each other. Let (x, y) be the strategy pair two label-$u\theta$ agents agree upon in a pairwise deviation. By the definition of an incomplete information blocking pair, $(x, y) \in NE_\theta$. We split into two cases:

- Suppose (x, y) is inefficient. Consider the type-θ_λ agent who agrees to play x facing a deviation plan $(D, \boldsymbol{x})$ such that $D = \{\theta, \tau\}$, $\boldsymbol{x}(\theta) = y$, and $\boldsymbol{x}(\tau) = x^+$.[31] Because $(x, y) \in NE_\theta$, we have $\pi(x, x^+) + \pi(x^+, x) \leqslant M^n$. In this case, the utility of the type-θ_λ

[31] Note that $x^+ \in X$ is indeed a rational and profitable play of a type-τ_λ agent facing a type-θ_λ partner in the deviation.

agent in the deviation is no more than

$$\begin{aligned} q_\lambda[\theta](M^n+\alpha)+(1-q_\lambda[\theta])M^n &= M^n+q_\lambda[\theta]\alpha \\ &\leqslant M^n+\frac{M-M^n}{\alpha}\alpha \\ &= M. \end{aligned}$$

Therefore, the agent in question does not benefit from playing x in the pairwise deviation against $(D,\boldsymbol{x})$.

- Suppose (x,y) is efficient and $\text{supp}(x)\subseteq X^+$ without loss. Consider the type-θ_λ agent who agrees to play x facing a deviation plan $(D,\boldsymbol{x})$ such that $D=\{\theta,\tau\}$, $\boldsymbol{x}(\theta)=y$, and $\boldsymbol{x}(\tau)=x^+$. In this case, the utility of the type-θ_λ agent in the deviation is no more than

$$\begin{aligned} q_\lambda[\theta](M+\alpha)+(1-q_\lambda[\theta])M^p &= q_\lambda[\theta](M-M^p+\alpha)+M^p \\ &\leqslant \frac{M-M^p}{M-M^p+\alpha}(M-M^p+\alpha)+M^p \\ &= M. \end{aligned}$$

Again, the agent in question does not benefit from playing x in the pairwise deviation against $(D,\boldsymbol{x})$.

Hence, the outcome (Λ,p,q,μ,S) is Bayes-Nash stable. It is left to verify that $G_\tau(\Lambda,p,q,\mu,S)>\frac{M}{2}>G_\theta(\Lambda,p,q,\mu,S)$. To see this, simply note that more than a half of type-τ agents play the advantageous strategy x^+ against y^-, while the opposite is true for type-θ agents. □

A.3 Proofs for Section 5.1: Polymorphism

A.3.1 Proof of Proposition 8

In this section, we first reformulate our model as a large roommate market in the language of Carmona and Laohakunakorn (2024). Next, we illustrate how a stable roommate matching in their setting can be transformed into a Nash stable outcome in our framework. Finally, we present their existence result and show that the conditions for existence are satisfied in the reformulated model.

Given the primitives of our model, define a **roommate market** $\mathcal{E}=(\Theta_\nu,\nu,C,\mathbb{C},(\succ_t)_{t\in\Theta_\nu})$ as follows. There is a finite set of types Θ_ν and a type distribution $\nu\in\Delta(\Theta_\nu)$. Let $\varnothing$ be a dummy type used to describe an unmatched agent, and write $\bar{\Theta}_\nu=\Theta_\nu\cup\{\varnothing\}$. There is a set

of **contracts** $C = \mathcal{X}^2$ and a **contract correspondence** $\mathbb{C} : \Theta_\nu \times \bar{\Theta}_\nu \rightrightarrows C$ such that

$$\mathbb{C}(t, t') = \begin{cases} NE_{t,t'} & \text{if } t' \neq \varnothing, \\ C & \text{if } t' = \varnothing, \end{cases}$$

where $NE_{t,t'}$ is the set of Nash equilibria between agents of types t and t'. Because Θ_ν is finite, we can fix an arbitrary order $\geqslant$ on types and require that $NE_{t,t'}$ specifies the strategy played by the bigger type as the first component. With this normalization, we have $\mathbb{C}(t, t') = \mathbb{C}(t', t)$ for all $t, t' \in \Theta_\nu$, satisfying the symmetry condition in Carmona and Laohakunakorn (2024). For each type $t \in \Theta_\nu$, let $\succ_t$ be a binary relation on $\bar{\Theta}_\nu \times C$ induced by the following utility function:

$$u_t^{\succ}(t', x, y) = \begin{cases} u_t(x, y, t') & \text{if } t' \in \Theta_\nu, \\ \underline{u} & \text{if } t' = \varnothing, \end{cases}$$

where $\underline{u} < \min_{t,t',x,y} u_t(x, y, t')$.

A **roommate matching** is a measure $\varphi \in \mathcal{M}(\Theta_\nu \times \bar{\Theta}_\nu \times C)$ such that $\operatorname{supp}(\varphi) \subseteq \operatorname{graph}(\mathbb{C})$ and $\nu_M + \nu_U + \nu_W = \nu$, where for each type $t \in \Theta_\nu$, $\nu_M[t] = \varphi[\{t\} \times \Theta_\nu \times C]$, $\nu_W[t] = \varphi[\Theta_\nu \times \{t\} \times C]$, and $\nu_U[t] = \varphi[\{t\} \times \{\varnothing\} \times C]$.[32] To define stability, we first describe the set of type-contract pairs that a particular type t can attract in a deviation, which we call type t's targets:

$$\begin{aligned} T_t(\varphi) = \big\{ &(t^*, x, y) \in \Theta_\nu \times C : (x, y) \in \mathbb{C}(t, t^*) \text{ and } \exists (t', x', y') \in \bar{\Theta}_\nu \times C \text{ s.t.} \\ &\operatorname{supp}(\varphi) \cap \{(t^*, t', x', y'), (t', t^*, x', y')\} \neq \varnothing \text{ and } (t, x, y) \succ_{t^*} (t', x', y') \big\}. \end{aligned}$$

Moreover, let $T_t^U(\varphi) = \{\varnothing\} \times C$ and $\bar{T}_t(\varphi) = T_t(\varphi) \cup T_t^U(\varphi)$. For a matching to be stable, no type-t agent can benefit from accepting some type-contract pair in $\bar{T}_t(\varphi)$. Thus, we write $S(\varphi)$ for the set of $(t, t', x, y) \in \Theta_\nu \times \bar{\Theta}_\nu \times C$ such that

(i) There does not exist $(\hat{t}, \hat{x}, \hat{y}) \in \bar{T}_t(\varphi)$ such that $(\hat{t}, \hat{x}, \hat{y}) \succ_t (t', x, y)$;

(ii) If $t' \neq \varnothing$, there does not exist $(\hat{t}, \hat{x}, \hat{y}) \in \bar{T}_{t'}(\varphi)$ such that $(\hat{t}, \hat{x}, \hat{y}) \succ_{t'} (t, x, y)$.

A roommate matching φ is **stable** if $\operatorname{supp}(\varphi) \subseteq S(\varphi)$.

The following lemma shows that we can always obtain a Nash stable outcome (μ, S) from a stable roommate matching φ. In fact, the two definitions are equivalent, but only one direction is important for our existence result.

Lemma 7. *Suppose there exists a stable roommate matching in the roommate market $\mathcal{E}$. Then there exists a Nash stable outcome under population distribution ν.*

[32] For a metric space Z, $\mathcal{M}(Z)$ is the set of finite Borel measures on Z.

Proof. Let φ be the stable roommate matching. We first show that $(\Theta_\nu \times \{\varnothing\} \times C) \cap \text{supp}(\varphi) = \varnothing$; that is, by construction, no agent is unmatched in φ. Towards a contradiction, suppose there exists a tuple $(t, \varnothing, x, y) \in (\Theta_\nu \times \{\varnothing\} \times C) \cap \text{supp}(\varphi)$. Fixing an arbitrary $(\widehat{x}, \widehat{y}) \in \mathbb{C}(t,t) \neq \varnothing$, we have $(t, \widehat{x}, \widehat{y}) \succ_t (\varnothing, x, y)$ by the definition of $\succ_t$. This means $(t, \widehat{x}, \widehat{y}) \in T_t(\varphi)$. On the other hand, note that $(t, \widehat{x}, \widehat{y}) \succ_t (\varnothing, x, y)$ also implies $(t, \varnothing, x, y) \notin S(\varphi)$. Thus, $\text{supp}(\varphi) \nsubseteq S(\varphi)$, contradicting the assumption that φ is a stable roommate matching.

Next, we construct an outcome (μ, S) in our setting from φ. For each type $t \in \Theta_\nu$, let

$$\mu_t[t'] = \frac{1}{\nu[t]} \left(\varphi[\{t\} \times \{t'\} \times C] + \varphi[\{t'\} \times \{t\} \times C]\right) \quad \text{for all } t' \in \Theta_\nu.$$

Since $\varphi[\{t\} \times \Theta_\nu \times C] + \varphi[\Theta_\nu \times \{t\} \times C] + \varphi[\{t\} \times \varnothing \times C] = \nu[t]$ and $\varphi[\Theta_\nu \times \varnothing \times C] = 0$, we have $\sum_{t' \in \Theta_\nu} \mu_t[t'] = 1$, meaning that $\mu_t \in \Delta(\Theta_\nu)$ is a well-defined probability distribution. The consistency condition $\nu[t]\mu_t[t'] = \nu[t']\mu_{t'}[t]$ is apparent from the definition. For each pair $t, t' \in \Theta_\nu$ such that $t \geqslant t'$ and $\mu_t[t'] > 0$, let

$$\begin{aligned} s_{t,t'}[E] &= \frac{1}{\nu[t]\mu_t[t']} \left(\varphi_f[\{t\} \times \{t'\} \times E] + \varphi_f[\{t'\} \times \{t\} \times E]\right), \text{ and} \\ s_{t',t}[E] &= s_{t,t'}[\rho(E)] \quad \text{for all } E \subseteq \mathcal{X}^2. \end{aligned}$$

If $\mu_t[t'] = 0$, then $s_{t,t'} \in \Delta_f(\mathcal{X}^2)$ can be defined arbitrarily. The strategy profile S is a vector that contains all $s_{t,t'}$.

Finally, we show that (μ, S) is Nash stable. Internal stability is implied by the fact that $\text{supp}(\varphi) \subseteq \text{graph}(\mathbb{C})$ and $\mathbb{C}(t,t') = NE_{t,t'}$ for every $t, t' \in \Theta_\nu$. To verify external stability, suppose by contradiction that a blocking pair exists. That is, there exist types $t, t', \bar{t}, \bar{t}' \in \Theta_\nu$ and strategy pairs $(\widehat{x}, \widehat{y}), (x', y'), (x'', y'') \in \mathcal{X}^2$ such that

(i) $\mu_t[\bar{t}] > 0$, $\mu_{t'}[\bar{t}'] > 0$, $(x', y') \in \text{supp}(s_{t,\bar{t}})$, and $(x'', y'') \in \text{supp}(s_{t',\bar{t}'})$;

(ii) $\widehat{x} \in \arg\max_{x \in \mathcal{X}} u_t(x, \widehat{y}, t')$ and $\widehat{y} \in \arg\max_{y \in \mathcal{X}} u_{t'}(y, \widehat{x}, t)$;

(iii) $u_t(\widehat{x}, \widehat{y}, t') > u_t(x', y', \bar{t})$ and $u_{t'}(\widehat{y}, \widehat{x}, t) > u_{t'}(x'', y'', \bar{t}')$.

$\mu_{t'}[\bar{t}'] > 0$, $(x'', y'') \in \text{supp}(s_{t',\bar{t}'})$, condition (ii), and $u_{t'}(\widehat{y}, \widehat{x}, t) > u_{t'}(x'', y'', \bar{t}')$ together imply that $(t', \widehat{x}, \widehat{y}) \in T_t(\varphi)$. Since $u_t(\widehat{x}, \widehat{y}, t') > u_t(x', y', \bar{t})$, we have $(t, \bar{t}, x', y') \notin S(\varphi_f)$. However, $\mu_t[\bar{t}] > 0$ and $(x', y') \in \text{supp}(s_{t,\bar{t}})$ imply that $(t, \bar{t}, x', y') \in \text{supp}(\varphi)$, which means $\text{supp}(\varphi) \nsubseteq S(\varphi)$. This contradicts the assumption that φ is stable. Therefore, (μ, S) is a Nash stable outcome. □

We say that the roommate market $\mathcal{E}$ is **acyclic** if $\succ_t$ is acyclic for each $t \in \Theta_\nu$.[33] Moreover,

[33] A relation $\succ$ on a set Z is acyclic if there is no finite sequence $\{z_1, z_2, \ldots, z_n\}$ such that $z_1 \succ z_2 \succ \cdots \succ z_n \succ z_1$.

$\mathcal{E}$ is **continuous** if $\{(t,c,t',c',t^*) \in (\bar{\Theta}_\nu \times C)^2 \times \Theta_\nu\} : (t,c) \succ_{t^*} (t',c')\}$ is open, $\mathbb{C}$ is continuous with nonempty and compact values, and $\Theta_\nu \times C$ is closed. We now state the existence result of Carmona and Laohakunakorn (2024).

Lemma 8. *If $\mathcal{E}$ is an acyclic and continuous roommate market, then $\mathcal{E}$ has a stable roommate matching.*

It remains to check that the roommate market $\mathcal{E}$ we defined is indeed acyclic and continuous. Because $\succ_t$ is induced by a utility function $u_t^{\succ}(t',x,y)$, $\succ_t$ is acyclic and $\{(t,c,t',c',t^*) \in (\bar{\Theta}_\nu \times C)^2 \times \Theta_\nu\} : (t,c) \succ_{t^*} (t',c')\}$ is open. Since Θ_ν is finite, $\mathbb{C}$ is continuous and $\Theta_\nu \times C$ is closed. For any $t, t' \in \Theta_\nu$, the set of Nash equilibria $NE_{t,t'}$ is nonempty and closed (and therefore compact). Hence, $\mathbb{C}$ has nonempty and compact values. Applying Lemmas 8 and 7 establishes the existence of a Nash stable outcome under population distribution ν.

A.3.2 Proof of Proposition 9

For part (i), suppose (μ, S) is Nash stable under ν and $G_\theta(\mu,S) > G_{\theta'}(\mu,S)$ for some $\theta, \theta' \in \Theta_\nu$. Then we must have $G_{\widehat{\theta}}(\mu,S) < \frac{M}{2}$ for some $\widehat{\theta} \in \Theta_\nu$; for if not, all types obtain an average material payoff weakly higher than $\frac{M}{2}$ while $G_\theta(\mu,S) > \frac{M}{2}$, which is impossible. Now suppose τ is parochial efficient and let $\widetilde{\nu} = (1-\varepsilon)\nu + \varepsilon\delta_\tau$ for some $\varepsilon \in (0, \bar{\varepsilon})$. Define an outcome $(\widetilde{\mu}, \widetilde{S})$ under $\widetilde{\nu}$ as follows. The matching profile $\widetilde{\mu} = (\widetilde{\mu}_t)$ is such that $\widetilde{\mu}_t = \mu_t$ for $t \in \Theta_\nu \backslash \{\tau\}$ and $\widetilde{\mu}_\tau[\tau] = 1$. Moreover, the strategy profile $\widetilde{S} = (\widetilde{s}_{t,t'})$ is such that $\widetilde{s}_{t,t'} = s_{t,t'}$ for $t, t' \in \Theta_\nu \backslash \{\tau\}$, $\widetilde{s}_{\tau,\tau}$ assigns probability one to efficient strategy pairs, and all other strategy distributions are arbitrary. It is easy to verify that $(\widetilde{\mu}, \widetilde{S})$ is a Nash stable outcome under $\widetilde{\nu}$. Moreover, we have $G_\tau(\widetilde{\mu}, \widetilde{S}) > G_{\widehat{\theta}}(\widetilde{\mu}, \widetilde{S})$, contradicting the assumption that ν is locally neutrally stable.

For part (ii), suppose $\mu_\theta[\theta'] > 0$, $(x,y) \in \text{supp}(s_{\theta,\theta'})$, and (x,y) is inefficient. This means $\sum_{t \in \Theta_\nu} \nu[\theta] G_\theta(\mu,S) < \frac{M}{2}$, which implies that $G_{\widehat{\theta}}(\mu,S) < \frac{M}{2}$ for some $\widehat{\theta} \in \Theta_\nu$. We can then follow the argument above and reach a contradiction. Note that this means $G_\theta(\mu,S) = \frac{M}{2}$ for all types $\theta \in \Theta_\nu$ if (μ,S) is Nash stable.

Now suppose (x,y) is efficient, but $\theta \neq \theta'$ and $\pi(x,y) > \pi(y,x)$. This means we must have $\pi(x',y') > \pi(y',x')$ for all $(x',y') \in \text{supp}(s_{\theta,\theta'})$. For if not, i.e. $\pi(x',y') \leqslant \pi(y',x')$ for some $(x',y') \in \text{supp}(s_{\theta,\theta'})$, we can redefine a strategy profile S' from S by replacing $s_{\theta,\theta'}$ with $s'_{\theta,\theta'}$ such that $s'_{\theta,\theta'}[(x',y')] = 1$; the outcome (μ, S') is also Nash stable but we have $G_\theta(\mu,S') < G_\theta(\mu,S) = \frac{M}{2}$, a contradiction. In order to ensure $G_{\theta'}(\mu,S) = \frac{M}{2}$, there must exist another $\theta'' \in \Theta_\nu$ such that $\mu_{\theta'}[\theta''] > 0$ and $\pi(x,y) > \pi(y,x)$ for all $(x,y) \in \text{supp}(s_{\theta',\theta''})$. Because Θ_ν is finite, we can repeat this argument and identify a set of types $\Theta_\nu^\circ = \{\theta_1, \theta_2, \ldots, \theta_k\} \subseteq \Theta_\nu$ such that the following holds: For each $1 \leqslant i \leqslant k$,

$\mu_{\theta_i}[\theta_{i+1}] > 0$ and $\pi(x,y) > \pi(y,x)$ for all $(x,y) \in \text{supp}(s_{\theta_i,\theta_{i+1}})$, with the interpretation that $k+1 = 1$. We now split into two cases:

- If k is even, then for $\zeta > 0$ sufficiently small, we can redefine μ' from μ as follows:

$$\mu'_{\theta_i}[\theta_{i+1}] = \mu_{\theta_i}[\theta_{i+1}] + \frac{\zeta}{\nu[\theta_i]},\ \mu'_{\theta_i}[\theta_{i-1}] = \mu_{\theta_i}[\theta_{i-1}] - \frac{\zeta}{\nu[\theta_i]} \quad \text{for each odd } i, \text{ and}$$
$$\mu'_{\theta_i}[\theta_{i+1}] = \mu_{\theta_i}[\theta_{i+1}] - \frac{\zeta}{\nu[\theta_i]},\ \mu'_{\theta_i}[\theta_{i-1}] = \mu_{\theta_i}[\theta_{i-1}] + \frac{\zeta}{\nu[\theta_i]} \quad \text{for each even } i.$$

  Other than these key components, all remaining parts are the same as in μ. It is easy to verify that μ' is a well-defined matching profile and (μ', S) is also Nash stable. However, the constructed outcome must satisfy $G_{\theta_i}(\mu', S) > G_{\theta_{i+1}}(\mu', S)$ for all odd i, a contradiction.

- If k is odd, then consider the invasion of a type $\tau = \theta_1$. Let $\varepsilon > 0$ be sufficiently small and consider the population distribution $\widetilde{\nu} = (1-\varepsilon)\nu + \varepsilon\delta_\tau$. Define a matching profile μ' under $\widetilde{\nu}$ as follows:

$$\mu'_{\theta_1}[\theta_2] = \frac{(1-\varepsilon)\nu[\theta_1]\mu_{\theta_1}[\theta_2] + \frac{\varepsilon}{2}}{\widetilde{\nu}[\theta_1]},\ \mu'_{\theta_1}[\theta_k] = \frac{(1-\varepsilon)\nu[\theta_1]\mu_{\theta_1}[\theta_k] + \frac{\varepsilon}{2}}{\widetilde{\nu}[\theta_1]},$$
$$\mu'_{\theta_i}[\theta_{i+1}] = \mu_{\theta_i}[\theta_{i+1}] - \frac{\varepsilon}{2\widetilde{\nu}[\theta_i]},\ \mu'_{\theta_i}[\theta_{i-1}] = \mu_{\theta_i}[\theta_{i-1}] + \frac{\varepsilon}{2\widetilde{\nu}[\theta_i]} \quad \text{for each even } i, \text{ and}$$
$$\mu'_{\theta_i}[\theta_{i+1}] = \mu_{\theta_i}[\theta_{i+1}] + \frac{\varepsilon}{2\widetilde{\nu}[\theta_i]},\ \mu'_{\theta_i}[\theta_{i-1}] = \mu_{\theta_i}[\theta_{i-1}] - \frac{\varepsilon}{2\widetilde{\nu}[\theta_i]} \quad \text{for each odd } i \neq 1.$$

  For all remaining components of μ' not specified above, let $\mu'_\theta[\cdot] = \frac{(1-\varepsilon)\nu[\theta]\mu_\theta[\cdot]}{\widetilde{\nu}[\theta]}$.[34] Again, one can verify that μ' is a well-defined matching profile and (μ', S) is a Nash stable outcome under the post-entry $\widetilde{\nu}$. However, we have $G_{\theta_1}(\mu', S) > G_{\theta_i}(\mu', S)$ for all even i, contradicting the assumption that ν is locally neutrally stable.

Finally, for part (iii), suppose $\pi(\widetilde{x}, \widetilde{y}) \neq \pi(\widetilde{y}, \widetilde{x})$ for all efficient strategy pairs $(\widetilde{x}, \widetilde{y})$. Then by part (ii), there cannot be a positive mass of cross-type matches, i.e. for every Nash stable outcome (μ, S) under ν, we have $\mu_\theta[\theta] = 1$ for all $\theta \in \Theta_\nu$. If θ exhibits same-type inefficiency, we can apply Lemma 1 and construct another Nash stable outcome (μ, S') such that $G_\theta(\mu, S') < \frac{M}{2}$, which contradicts the conclusion in part (ii). If $u_\theta(x, y, t)$ is constant in t, we can instead follow the proof of Proposition 2 and construct a mutant type τ that dominates θ in the post-entry population. Therefore, ν cannot be locally neutrally stable, a contradiction.

[34] Note that $\widetilde{\nu}[\theta_1] = (1-\varepsilon)\nu[\theta_1] + \varepsilon$ and $\widetilde{\nu}[\theta] = (1-\varepsilon)\nu[\theta]$ for all $\theta \neq \theta_1$.

A.3.3 Proof of Proposition 10

The proof is a straightforward extension of that of Proposition 1. Suppose Θ_ν consists of homophilic and/or parochial efficient types, and consider any mutant type $\tau \in \Theta$ and $\varepsilon \in (0,1)$. For any Nash stable outcome $(\widetilde{\mu}, \widetilde{S})$ under the post-entry population $\widetilde{\nu} = (1-\varepsilon)\nu + \varepsilon\delta_\tau$, we must have $\widetilde{\mu}_\theta[\theta] = 1$ for all $\theta \in \Theta_\nu$. For if not, two type-θ agents who are matched with another type can form a blocking pair by coordinating on the efficient strategy pair with each other. This also implies that $\widetilde{\mu}_\tau[\tau] = 1$. Now in the spirit of Lemma 1, the strategy distribution $s_{\theta,\theta}$ must attach probability one to efficient strategy pairs, meaning that $G_\theta(\widetilde{\mu}, \widetilde{S}) \geqslant G_\tau(\widetilde{\mu}, \widetilde{S})$ for all $\theta \in \Theta_\nu$. Therefore, ν is locally neutrally stable.

References

Ahn, T. K., R. Mark Isaac, and Timothy C. Salmon (2009): "Coming and going: Experiments on endogenous group sizes for excludable public goods," *Journal of Public Economics*, 93, 336–351.

Aimone, Jason A., Laurence R. Iannaccone, Michael D. Makowsky, and Jared Rubin (2013): "Endogenous Group Formation via Unproductive Costs," *The Review of Economic Studies*, 80 (4), 1215–1236.

Akçay, E., J. Van Cleve, M. Feldman, and J. Roughgarden (2009): "A Theory for the Evolution of Other-regard Integrating Proximate and Ultimate Perspectives," *Proceedings of the National Academy of Sciences*, 106 (45), 19061–19066.

Akdeniz, Aslihan and Matthijs van Veelen (2021): "The evolution of morality and the role of commitment," *Evolutionary Human Sciences*, 3, e41.

Akdeniz, Aslihan and Matthijs van Veelen (2023): "Evolution and the ultimatum game," *Games and Economic Behavior*, 142, 570–612.

Alger, I. (2010): "Public Goods Games, Altruism, and Evolution," *Journal of Public Economic Theory*, 12 (4), 789–813.

Alger, Ingela (2022): "Evolutionarily Stable Preferences," *Philosophical Transactions of the Royal Society B*, 378, 20210505.

Alger, Ingela and Laurent Lehmann (2023): "Evolution of semi-Kantian preferences in two-player assortative interactions with complete and incomplete information and plasticity," *Dynamic games and Applications*, forthcoming.

Alger, I. and J. W. Weibull (2010): "Kinship, Incentives, and Evolution," *American*

Economic Review, 100 (4), 1725–1758.

Alger, Ingela and Jörgen W. Weibull (2012): "A Generalization of Hamilton's Rule - Love The Sibling How Much?" *Journal of Theoretical Biology*, 299, 42–54.

——— (2013): "Homo Moralis: Preference Evolution under Incomplete Information and Assortative Matching," *Econometrica*, 81 (6), 2269–2302.

——— (2016): "Evolution and Kantian morality," *Games and Economic Behavior*, 98, 56–67.

——— (2019): "Evolutionary Models of Preference Formation," *Annual Review of Economics*, 11, 329–354.

Alger, Ingela, Jörgen W. Weibull, and Laurent Lehmann (2020): "Evolution of preferences in structured populations: Genes, guns, and culture?" *Journal of Economic Theory*, 185, 10495.

Ali, S. Nageeb and Ce Liu (2025): "Coalitions in Repeated Games," Working paper.

Avataneo, Michelle, Nicola Persico, and Thomas Norman (2025): "The Evolutionary Stability of Moral Foundations," Working paper.

Bandhu, Sarvesh and Ratul Lahkar (2023): "Survival of altruistic preferences in a large population public goods game," *Economics Letters*, 226, 111113.

Baumard, Nicolas, Jean-Baptiste André, and Dan Sperber (2013): "A mutualistic approach to morality: The evolution of fairness by partner choice," *Behavioral and Brain Sciences*, 36, 59–122.

Becker, Gary S. (1976): "Altruism, Egoism, and Genetic Fitness," *Journal of Economic Literature*, 14, 817–826.

Bernhard, H., U. Fischbacher, and E. Fehr (2006): "Parochial altruism in humans," *Nature*, 442, 912–915.

Bester, H. and W. Güth (1998): "Is Altruism Evolutionarily Stable," *Journal of Economic Behavior and Organization*, 34, 193–209.

Bilancini, E. and L. Boncinelli (2018): "Social coordination with locally observable types," *Economic Theory*, 65, 975–1009.

Bisin, Alberto, Jared Rubin, Avner Seror, and Thierry Verdier (2021): "Culture, Institutions & the Long Divergence," *NBER Working Papers 28488 National Bureau of Economic Research, Inc.*

Bisin, Alberto and Thierry Verdier (2021): "Phase Diagrams in Historical Economics: Culture and Institutions," in *Handbook of Historical Economics*, ed. by A. Bisin and

G. Federico, Amsterdam: Elsevier North Holland.

——— (2024): "On the Joint Evolution of Culture and Political Institutions: Elites and Civil Society," *Journal of Political Economy*, 132 (5), 1485–1564.

Bolle, F. (2000): "Is altruism evolutionarily stable? And envy and malevolence?: Remarks on Bester and Güth," *Journal of Economic Behavior & Organization*, 42, 131–133.

Brekke, Kjell Arne, Karen Evelyn Hauge, Jo Thori Lind, and Karine Nyborg (2011): "Playing with the good guys: A public good game with endogenous group formation," *Journal of Public Economics*, 95, 1111–1118.

Burlando, Roberto M. and Francesco Guala (2005): "Heterogeneous Agents in Public Goods Experiments," *Experimental Economics*, 8, 35–54.

Carmona, Guilherme and Krittanai Laohakunakorn (2024): "Stable Matching in Large Markets with Occupational Choice," *Theoretical Economics*, 19, 1261–1304.

Carvalho, Jean-Paul, Augustin Bergeron, Joseph Henrich, Nathan Nunn, and Jonathan Weigel (2023): "Zero-Sum Thinking, the Evolution of Effort-Suppressing Beliefs, and Economic Development," *Working paper*.

Charness, G. and M. Rabin (2002): "Understanding Social Preferences with Simple Tests," *Quarterly Journal of Economics*, 117, 817–869.

Charness, Gary and Chun-Lei Yang (2014): "Starting small: Toward voluntary formation of efficient large groups in public goods provision," *Journal of Economic Behavior & Organization*, 102, 119–132.

Chen, Daniel L. and Martin Schonger (2022): "Social preferences or sacred values? Theory and evidence of deontological motivations," *Science Advances*, 8 (19).

Chen, Yi-Chun and Gaoji Hu (2023): "A Theory of Stability in Matching with Incomplete Information," *American Economic Journal: Microeconomics*, 15 (1), 288–322.

Choi, J. and S. Bowles (2007): "The Coevolution of Parochial Altruism and War," *Science*, 318, 636–640.

Coricelli, Giorgio, Dietmar Fehr, and Gerlinde Fellner (2004): "Partner Selection in Public Goods Experiments," *The Journal of Conflict Resolution*, 48 (3), 356–378.

Cui, Z. and F. Shi (2021): "Bandwagon Effects and Constrained Network Formation," *Games and Economic Behavior*, 134, 37–51.

Cui, Z. and S. Weidenholzer (2021): "Lock-in Through Passive Connections," *Journal of Economic Theory*, 192, 105187.

de Oliveira, Angela C. M., Rachel T. A. Croson, and Catherine Eckel (2015): "One bad apple? Heterogeneity and information in public good provision," *Experimental Economics*, 18, 116–135.

Dekel, E., J. C. Ely, and O. Yilankaya (2007): "Evolution of Preferences," *Review of Economics Studies*, 74, 685–704.

Echenique, Federico, Sangmok Lee, Matthew Shum, and Bumin M. Yenmez (2013): "The Revealed Preference Theory of Stable and Extremal Stable Mathcings," *Econometrica*, 81 (1), 153–171.

Ehrhart, Karl-Martin and Claudia Keser (1999): "Mobility and Cooperation: On the Run," *Cirano Working paper*.

Ellingsen, T. (1997): "The Evolution of Bargaining Behavior," *The Quarterly Journal of Economics*, 112 (2), 581–602.

Ely, Jeffrey C. (2002): "Local Conventions," *The B.E. Journal of Theoretical Economic*, 149, 1–32.

Ely, J. C. and O. Yilankaya (2001): "Nash Equilibrium and the Evolution of Preferences," *Journal of Economic Theory*, 97, 255–272.

Engelmann, Dirk and Martin Strobel (2004): "Inequality aversion, efficiency, and maximin preferences in simple distribution experiments," *American Economic Review*, 94 (4), 857–869.

Fershtman, C. and Y. Weiss (1998): "Social Rewards, Externalities and Stable Preferences," *Journal of Public Economics*, 70, 53–73.

Frank, Robert H. (1987): "If Homo Economicus Could Choose His Own Utility Function, Would He Want One with a Conscience?" *American Economic Review*, 77, 593–604.

——— (1988): *Passions Within Reason: The Strategic Role of the Emotions*, New York: Norton.

Fujiwara-Greve, Takako and Masahiro Okuno-Fujiwara (2009): "Voluntarily Separable Repeated Prisoner's Dilemma," *Review of Economic Studies*, 76, 993–1021.

Gächter, Simon and Christian Thöni (2005): "Social Learning and Voluntary Cooperation Among Like-Minded People," *Journal of the European Economic Association*, 3 (2-3), 303–314.

Gale, David and Lloyd S. Shapley (1962): "College Admissions and the Stability of Marriage," *The American Mathematical Monthly*, 69 (1), 9–15.

Garrido-Lucero, Felipe and Rida Laraki (2021): "Stable Matching Games," *Working paper*.

Gintis, H., E. Alden Smith, and S. Bowles (2001): "Costly Signaling and Cooperation," *Journal of Theoretical Biology*, 213, 103–119.

Goyal, S. and F. Vega-Redondo (2005): "Network formation and social coordination," *Games and Economic Behavior*, 50, 178–207.

Graser, Christopher, Takako Fujiwara-Greve, Julian García, and Matthijs van Veelen (2024): "Repeated Games with Partner Choice," Working paper.

Grimm, Veronika and Friederike Mengel (2009): "Cooperation in viscous populations—Experimental evidence," *Games and Economic Behavior*, 66, 202–220.

Guido, Andrea, Andrea Robbett, and Rustam Romaniuc (2019): "Group formation and cooperation in social dilemmas: A survey and meta-analytic evidence," *Journal of Economic Behavior and Organization*, 159, 192–209.

Gul, Faruk and Wolfgang Pesendorfer (2016): "Interdependent Preference Models as a Theory of Intentions," .

Gunnthorsdottir, Anna, Roumen Vragov, Stefan Seifert, and Kevin McCabe (2010): "Near-efficient equilibria in contribution-based competitive grouping," *Journal of Public Economics*, 94, 987–994.

Gürerk, Özgür, Bernd Irlenbusch, and Bettina Rockenbach (2014): "On cooperation in open communities," *Journal of Public Economics*, 120, 220–230.

Güth, W. (1995): "An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Incentives," *International Journal of Game Theory*, 24, 323–344.

Güth, Werner and Hartmut Kliemt (1998): "Indirect Evolutionary Approach: Bridging the Gap between Rationality and Adaptation," *Rationality and Society*, 10 (3), 377–399.

Güth, W. and M. Yaari (1992): "An Evolutionary Approach to Explain Reciprocal Behavior in a Simple Strategic Game," in *Explaining Process and Change - Approaches to Evolutionary Economics*, ed. by U. Witt, University of Michigan Press, 23–34.

Hamilton, W. D. (1964a): "The Genetical Evolution of Social Behaviour. I," *Journal of Theoretical Biology*, 7, 1–16.

——— (1964b): "The Genetical Evolution of Social Behaviour. II," *Journal of Theoretical Biology*, 7, 17–52.

Hauk, Esther and Rosemarie Nagel (2001): "Choice of Partners in Multiple Two-Person

Prisoner's Dilemma Games: An Experimental Study," *Journal of Conflict Resolution*, 45, 770–793.

Heifetz, A., C. Shannon, and Y. Spiegel (2007a): "The Dynamic Evolution of Preferences," *Economic Theory*, 32, 251–286.

——— (2007b): "What to Maximize if You Must," *Journal of Economic Theory*, 133, 31–57.

Heller, Yuval and Erik Mohlin (2019): "Coevolution of deception and preferences: Darwin and Nash meet Machiavelli," *Games and Economic Behavior*, 113, 223–247.

Herold, F. and C. Kuzmics (2009): "Evolutionary Stability of Discrimination under Observability," *Games and Economic Behavior*, 67, 542–551.

Hiller, Victor and Nouhoum Touré (2021): "Endogenous gender power: The two facets of empowerment," *Journal of Development Economics*, 149, 102596.

Hirshleifer, J. (1977): "Economics from a Biological Viewpoint," *Journal of Law and Economics*, 20, 1–52.

Hirshleifer, Jack (2001): "Game-theoretic interpretations of commitment," in *Evolution and the Capacity for Commitment*, ed. by Randolph M Nesse, New York: Russell Sage Foundation, 77–93.

Hojman, D. and A. Szeidil (2006): "Endogenous networks, social games, and evolution," *Games and Economic Behavior*, 55, 112–130.

Hopkins, Ed (2014): "Competitive Altruism, Mentalizing and Signaling," *American Economic Journal: Microeconomics*, 272–292.

Huck, S. and J. Oechssler (1999): "The Indirect Evolutionary Approach to Explaining Fair Allocations," *Games and Economic Behavior*, 28, 13–24.

Izquierdo, S. S., L. R. Izquierdo, and M. van Veelen (2021): "Repeated Games with Endogenous Separation," Working paper.

Izquierdo, Segismundo S., Luis R. Izquierdo, and Fernando Vega-Redondo (2010): "The Option to Leave: Conditional Dissociation in the Evolution of Cooperation," *Journal of Theoretical Biology*, 267, 76–84.

——— (2014): "Leave and let leave: A sufficient condition to explain the evolutionary emergence of cooperation," *Journal of Economic Dynamics and Control*, 46, 91–113.

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

Jackson, M. O. and A. Watts (2002): "On the formation of interaction networks in

social coordination games," *Games and Economic Behavior*, 41, 265–291.

——— (2010): "Social Games: Matching and the Play of Finitely Repeated Games," *Games and Economic Behavior*, 70, 170–191.

Jansen, M. J. M. (1981): "Maximal Nash Subsets for Bimatrix Games," *Naval Research Logistics Quarterly*, 28 (1), 147–152.

Kant, I. (1785): *Groundwork of the Metaphysics of Morals*, New York: Harper Torch Books.

Koçkesen, L., E. A. Ok, and R. Sethi (2000): "The Strategic Advantage of Negatively Interdependent Preferences," *Journal of Economic Theory*, 92, 274–299.

Koh, Paul S. (2023): "Stable Outcomes and Information in Games: An Empirical Framework," *Journal of Econometrics*, 237 (1), 105499.

Lahkar, Ratul (2019): "Elimination of Non-Individualistic Preferences in Large Population Aggregative Games," *Journal of Mathematical Economics*, 845, 150–165.

Lehmann, Laurent, Ingela Alger, and Jörgen W. Weibull (2015): "Does evolution lead to maximizing behavior?" *Evolution*, 69, 1858–1873.

Liu, Qingmin (2020): "Stability and Bayesian Consistency in Two-Sided Markets," *American Economic Review*, 110 (8), 2625–2666.

Liu, Qingmin, Mailath George J, Andrew Postlewaite, and Larry Samuelson (2014): "Stable Matching with Incomplete Information," *Econometrica*, 82 (2), 541–587.

Mailath, G., L. Samuelson, and A. Shaked (1997): "Endogenous interactions," *Working paper*.

McNamara, J. M., Z. Barta, L. Fromhage, and A. I. Houston (2008): "The coevolution of choosiness and cooperation," *Nature*, 451, 189–192.

McNamara, J. M., C. E. Gasson, and A. I. Houston (1999): "Incorporating rules for responding into evolutionary games," *Nature*, 401, 368–371.

Mill, J. S. (1863): *Utilitarianism*, Chicago: University of Chicago Press.

Nax, H. H. and A. Rigos (2016): "Assortativity Evolving from Social Dilemmas," *Journal of Theoretical Biology*, 395, 194–203.

Newton, Jonathan (2017): "The preferences of Homo Moralis are unstable under evolving assortativity," *International Journal of Game Theory*, 46, 583–589.

Ockenfels, P. (1993): "Cooperation in prisoners' dilemma: An evolutionary approach," *European Journal of Political Economy*, 9 (4), 567–579.

Ok, E. A. and F. Vega-Redondo (2001): "On the Evolution of Individualistic Preferences: An Incomplete Information Scenario," *Journal of Economic Theory*, 97, 231–254.

Page, Talbot, Louis Putterman, and Bulent Unel (2005): "Voluntary Association in Public Goods Experiments: Reciprocity, Mimicry and Efficiency," *The Economic Journal*, 115 (506), 1032–1053.

Possajennikov, A. (2000): "On the Evolutionary Stability of Altruistic and Spiteful Preferences," *Journal of Economic Behavior & Organization*, 42 (2), 125–129.

Rand, David G., Samuel Arbesman, and Nicholas A. Christakis (2011): "Dynamic social networks promote cooperation in experiments with humans," *Proceedings of the National Academy of Sciences*, 108 (48), 19193–19198.

Riedl, Arno, Ingrid M. T. Rohde, and Martin Strubel (2016): "Efficient Coordination in Weakest-Link Games," *The Review of Economic Studies*, 83 (2), 737–767.

Robson, Arthur and Larry Samuelson (2011): "The evolutionary foundations of preferences," in *Handbook of social economics*, ed. by J. Benhabib, A. Bisin, and M. Jackson, Elsevier, vol. 1, 221–310.

Robson, A. J. (1990): "Efficiency in Evolutionary Games: Darwin, Nash and the Secret Handshake," *Journal of Theoretical Biology*, 144, 379–396.

Rubin, P. and C. Paul (1979): "An Evolutionary Model of Taste for Risk," *Economic Inquiry*, 42, 585–596.

Sethi, R. and E. Somanathan (2001): "Preference Evolution and Reciprocity," *Journal of Economic Theory*, 97, 273–297.

Slonim, Robert and Ellen Garbarino (2008): "Increases in trust and altruism from partner selection: Experimental evidence," *Experimental Economics*, 11, 134–153.

Staudigl, M. and S. Weidenholzer (2014): "Constrained interactions and social coordination," *Journal of Economic Theory*, 152, 41–63.

Tomasello, M. (2016): *A Natural History of Human Morality*, Cambridge, Massachusetts: Harvard University Press.

Van Veelen, M. (2006): "Why kin and group selection models may not be enough to explain human other-regarding behaviour," *Journal of Theoretical Biology*, 242, 790–797.

Wang, Ziwei (2022): "Rationalizable Stability in Matching with One-Sided Incomplete Information," *Working paper*.

Weibull, J. W. (1995): *Evolutionary Game Theory*, Cambridge, MA: The MIT Press.

Wilson, David Sloan and Lee A. Dugatkin (1997): "Group Selection and Assortative Interactions," *The American Naturalist*, 149, 336–351.

Wu, Jiabin (2017): "Political institutions and the evolution of character traits," *Games and Economic Behavior*, 106, 260–276.

——— (2019): "Labelling, Homophily and Preference Evolution," *International Journal of Game Theory*, 49, 1–22.

O Online Appendix for "Preference Evolution under Stable Matching"

O.1 Weak Blocking

Consider an underlying game where the material payoffs are given by the following table ,

	A	B
A	$0,0$	$2,3$
B	$3,2$	$0,0$

We first assume all agents in the population are selfish. That is, their utility function coincides with the material payoffs in the table. According to our definition of blocking and stable outcomee, an outcome where all agents are matched to play (A,B) or (B,A) in each pair constitutes a (unique) stable outcome. Those agents receiving a payoff of 2 are the "losers" of the game, and they might seek to rematch. Therefore, if weak improvement is allowed for blocking, the outcome described above would no longer be considered stable.

However, adopting this weaker definition of blocking presents a problem: the existence of a stable outcome is no longer guaranteed (Jackson and Watts (2010) make a similar observation). To see this, first note that in a stable outcome, there cannot be a positive mass of matched agents playing the asymmetric equilibrium (A,B). For if so, there must be a positive mass of "losers" receiving a utility of 2; then any pair of such agents can form a pairwise deviation to the equilibrium (A,B), making one of them strictly better off. Therefore, the only possible stable outcome is one in which all agents are matched to play the symmetric equilibrium. However, in this scenario, any two agents can form a blocking pair and play the equilibrium (A,B) which makes both strictly better off. As a result, no stable outcome exists under the weaker definition of blocking. In contrast, if we require both inequalities to be strict in the definition of blocking, a stable outcome always exists, as we show in Proposition 8 in Appendix A.3.1.

		type-θ	
		A	B
type-θ	A	$0,0$	$5,5$
	B	$5,5$	$0,0$

	type-τ	
	A	B
A	$0,-$	$0,-$
B	$0,-$	$0,-$

Nevertheless, even if we allow weak improvement for a blocking pair, the main implications of our paper remain unchanged. We continue to use the material game in the table provided at the beginning of this appendix for illustration. When type-θ agents are parochial efficient, the stable outcome must exhibit perfectly assortative matching and efficient play between

type-θ agents, which lead to the evolutionary stability of parochial efficient preferences. The main rationales are two-fold:

- When two θ-agents are matched, they must play the asymmetric equilibrium (A, B) or (B, A). Moreover, neither agent is a "loser" because both of them derive a *utility* of $3 + 2 = 5$ (see the left table in the above set of tables);
- Parochialism ensures that type-θ agents have no incentive, not even a weak one, to match with agents of other preferences (see the right table in above set of tables). Therefore, regardless of how the type-τ agents behave (these agents may never settle due to the non-existence issue explained above), the type-θ agents obtain weakly higher average material payoffs than them.

O.2 Homophilic Efficient Types are Not Neutrally Stable under Incomplete Information

In this section, we provide a condition on the material game under which the homophilic efficient types are not neutrally stable in the case of incomplete information. This condition covers a wide class of games (e.g. the ones in Examples 2 and 4) and is different from the condition in Proposition 5. It therefore strengthens our conclusion that the homophilic efficient preferences are not favored by evolutionary forces.

Proposition 11. *If some inefficient strategy pair is a strict Nash equilibrium between two agents with efficient preferences, then any homophilic efficient type is not neutrally stable.*

Proof. Fixing $\alpha > 0$, denote by θ the α-homophilic efficient type. Suppose $(\widetilde{x}, \widetilde{y})$ is inefficient and is a strict Nash equilibrium in NE_θ. Write E for the set of efficient strategy pairs and S for the efficient total material payoffs,

$$S = \max_{(x,y)\in X^2} \pi(x, y) + \pi(y, x).$$

Moreover, let $\widetilde{S}$ be the total material payoffs when $(\widetilde{x}, \widetilde{y})$ is played and $\widehat{S}$ be the second highest total material payoffs that result in pure strategies when one agent plays $\widetilde{x}$, i.e.

$$\widetilde{S} = \pi(\widetilde{x}, \widetilde{y}) + \pi(\widetilde{y}, \widetilde{x}) \quad \text{and} \quad \widehat{S} = \max_{y\in X\setminus\{\widetilde{y}\}} \pi(\widetilde{x}, y) + \pi(y, \widetilde{x}).$$

By assumptions, we have $S > \widetilde{S} > \widehat{S}$.

Our goal is to construct a population state $(\theta, \tau, \varepsilon)$ and a Bayes-Nash stable outcome (p, μ, S) such that $G_\tau(p, \mu, S) > G_\theta(p, \mu, S)$. To this end, suppose $1 - \varepsilon < \frac{\widetilde{S}-\widehat{S}}{S-\widehat{S}}$ and consider

a preference type τ whose utility function is given by

$$u_\tau(x,y,t) = \begin{cases} \pi(x,y) + \pi(y,x) & \text{if } t = \tau, \\ \frac{S}{1-\varepsilon} \cdot \mathbb{1}_{\{x \in \{\widetilde{x},\widetilde{y}\}\}} & \text{otherwise.} \end{cases}$$

Now consider a matching profile (p,μ) that satisfies $p_\theta = 0$, $p_u, p_\tau > 0$, $q_{u\theta} \in \left(1-\varepsilon, \frac{\widetilde{S}-\widehat{S}}{S-\widehat{S}}\right]$, and $\mu_{u,u} = \mu_{\tau,\tau} = 1$. The strategy profile is given by $S = \{(\widetilde{x},\widetilde{y})_{u,u}, (x^*,y^*)_{\tau,\tau}\}$ where (x^*,y^*) is an arbitrary efficient strategy pair. This outcome is depicted in Figure 4 below.

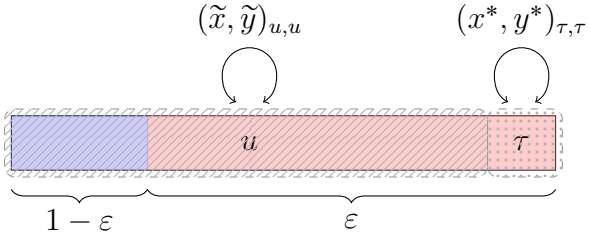

Figure 4: The matching profile for the proof of part (i).

We argue that (p,μ,S) is a Bayes-Nash stable outcome. First, because a type-τ agent always plays $\widetilde{x}$ or $\widetilde{y}$ against a type-θ agent, no label-$u\theta$ agent has an incentive to rematch with (or target) a type-τ agent. Second, because $\frac{S}{1-\varepsilon} \cdot q_{u\theta} > S$, no label-$u\tau$ agent has an incentive to rematch with (or target) a type-τ agent. Moreover, the strategy pair is efficient between label-$u\tau$ agents, so there is no blocking pair among them. Therefore, it is only left to consider blocking pairs consisting of two label-$u\theta$ agents who target each other. Let $(\breve{x},\breve{y})$ be the strategy pair two label-$u\theta$ agents agree upon in a pairwise deviation. By the definition of an incomplete information blocking pair, $(\breve{x},\breve{y}) \in NE_\theta$. Consider the label-$u\theta$ agent who agrees to play $\breve{y}$ facing a deviation plan $(D,\boldsymbol{x})$ such that $D = \{\theta,\tau\}$, $\boldsymbol{x}(\theta) = \breve{x}$, and $\boldsymbol{x}(\tau) = \widetilde{x}$. There are two cases to check:

- $\breve{y}$ attaches positive probability to $\widetilde{y}$, which implies $\pi(\breve{x},\breve{y}) + \pi(\breve{y},\breve{x}) = \pi(\breve{x},\widetilde{y}) + \pi(\widetilde{y},\breve{x})$ because type θ maximizes total material payoffs. Because $(\widetilde{x},\widetilde{y}) \in NE_\theta$, we must have $\pi(\breve{x},\widetilde{y}) + \pi(\widetilde{y},\breve{x}) \leqslant \widetilde{S}$. These two together imply that $\pi(\breve{x},\breve{y}) + \pi(\breve{y},\breve{x}) \leqslant \widetilde{S}$. Since $(\widetilde{x},\widetilde{y}) \in NE_\theta$, we also have $\pi(\breve{y},\widetilde{x}) + \pi(\widetilde{x},\breve{y}) \leqslant \widetilde{S}$. The deviation then yields no more than

  $$q_{u\theta}(\widetilde{S}+\alpha) + (1-q_{u\theta})\widetilde{S} = \widetilde{S} + q_{u\theta}\alpha,$$

  which is the utility of a label-$u\theta$ agent in the status quo.

- $\breve{y}$ does not attach positive probability to $\widetilde{y}$. The deviation then yields no more than

  $$q_{u\theta}(S+\alpha) + (1-q_{u\theta})\widehat{S} = q_{u\theta}(S-\widehat{S}) + q_{u\theta}\alpha + \widehat{S}$$

$$\leqslant \frac{\widetilde{S}-\widehat{S}}{S-\widehat{S}}(S-\widehat{S})+q_{u\theta}\alpha+\widehat{S}$$
$$=\widetilde{S}+q_{u\theta}\alpha.$$

Therefore, the label-$u\theta$ agent in question does not benefit from playing $\breve{y}$ in the pairwise deviation against $(D,\boldsymbol{x})$.

Hence, the outcome (p,μ,S) is Bayes-Nash stable. Observe that $G_\tau(p,\mu,S)>G_\theta(p,\mu,S)$ because $(\widetilde{x},\widetilde{y})$ is an inefficient strategy pair. We can conclude that the α-homophilic efficient type is not neutrally stable. □

In the construction of the population state and the outcome above, the lower bound on the proportion of type τ is crucial. Therefore, the proposition should be interpreted with caution: Any homophilic efficient type is unable to break from an inefficient outcome as the *invading minority* in a population, which leads to an average material payoff strictly lower than that of the incumbent. Therefore, the homophilic efficient types are not neutrally stable.

O.3 An Extension of Proposition 2 under Incomplete Information

Proposition 12. *With the presence of incomplete information,*

(i) *If θ exhibits same-type inefficiency, then θ is evolutionarily unstable;*

(ii) *If $\pi(\widetilde{x},\widetilde{y})\neq\pi(\widetilde{y},\widetilde{x})$ for any efficient strategy pair $(\widetilde{x},\widetilde{y})$ and $u_\theta(x,y,t)=f(x,y)$, then θ is evolutionarily unstable.*

Proof. (i) By Proposition 6, if θ exhibits same-type inefficiency, then the parochial efficient type is evolutionarily stable against θ. Thus, θ is evolutionarily unstable by definition.

(ii) Let θ be a type such that $u_\theta(x,y,t)=f(x,y)$. If θ exhibits same-type inefficiency, then part (i) applies. Thus, we suppose all strategy pairs in NE_θ^{lb} are efficient. Now consider another type τ that has the following utility function:

$$u_\tau(x,y,t)=\begin{cases}\pi(x,y)+\pi(y,x) & \text{if } t=\tau,\\ [\pi(x,y)+\pi(y,x)]\cdot\mathbb{1}_{\{\pi(x,y)\geqslant\pi(y,x)\}}\cdot\mathbb{1}_{\{(x,y)\text{ is efficient}\}} & \text{if } t=\theta.\end{cases}$$

We now show that τ is evolutionarily stable against θ.

First, we argue that $G_\tau(p,\mu,S)\geqslant G_\theta(p,\mu,S)$ for all Bayes-Nash stable outcomes. This is because by construction of type τ's utility function, the following properties hold for any Bayes-Nash stable outcome:

- If $p_\tau>0$ and $(x^*,y^*)_{\tau,\tau}\in S$, then (x^*,y^*) is efficient;

- If $p_\tau, p_\theta > 0$ and $(x^*, y^*)_{\tau,\theta} \in S$, then (x^*, y^*) is efficient and $\pi(x^*, y^*) \geqslant \pi(y^*, x^*)$;
- If $p_u\mu_{u,\tau} > 0$ and $(x^*, y^*)_{u,\tau} \in S$, then (x^*, y^*) is efficient and $\pi(y^*, x^*) \geqslant \pi(x^*, y^*)$;
- If $p_u\mu_{u,u} > 0$ and $(x^*, y^*)_{u,u} \in S$, then (x^*, y^*) is efficient;
- If $p_u\mu_{u,\theta} > 0$ and $(x^*, y^*)_{u,\theta} \in S$, then (x^*, y^*) is efficient and $\pi(x^*, y^*) \geqslant \pi(y^*, x^*)$.

When the first three properties do not hold, two label-τ agents can form a blocking pair and coordinate on any efficient strategy pair. When the last two properties are violated, two label-$u\tau$ agents can form a blocking pair with strong incentives. The argument for the validity of blocking pairs is similar to that in the proof of Proposition 6 and thus omitted. Write $(\widetilde{x}, \widetilde{y})$ for an arbitrary efficient strategy pair. We then have

$$G_\tau(p, \mu, S) \geqslant \frac{1}{2}\pi(\widetilde{x}, \widetilde{y}) + \frac{1}{2}(\widetilde{y}, \widetilde{x}) \geqslant G_\theta(p, \mu, S).$$

because type-τ agents perform, on average, weakly better than type-θ agents in all possible matches. (In particular, if $p_u\mu_{u,\theta} > 0$, some type-θ agents (label-$u\theta$ agents) are playing an advantageous strategy, but the *average* payoff of type θ is lower than $\frac{1}{2}\pi(\widetilde{x}, \widetilde{y}) + \frac{1}{2}(\widetilde{y}, \widetilde{x})$.)

It remains to show that the inequality is strict for some (a large set of) Bayes-Nash stable outcomes. Recall that we assume $\pi(\widetilde{x}, \widetilde{y}) \neq \pi(\widetilde{y}, \widetilde{x})$ for all efficient strategy pairs $(\widetilde{x}, \widetilde{y})$, so we have the following three possibilities

(i) $p_\tau, p_\theta > 0$, $(\widetilde{x}, \widetilde{y})_{\tau,\theta} \in S$, $(\widetilde{x}, \widetilde{y}) \in NE^{lb}_\theta$, and $\pi(\widetilde{x}, \widetilde{y}) > \pi(\widetilde{y}, \widetilde{x})$;

(ii) $p_u\mu_{u,\tau} > 0$, $(\widetilde{y}, \widetilde{x})_{u,\tau} \in S$, $(\widetilde{y}, \widetilde{x}) \in NE^{lb}_\theta$, and $\pi(\widetilde{x}, \widetilde{y}) > \pi(\widetilde{y}, \widetilde{x})$;

(iii) $p_u\mu_{u,\theta} > 0$, $(\widetilde{x}, \widetilde{y})_{u,\theta} \in S$, $(\widetilde{x}, \widetilde{y}) \in NE^{lb}_\theta$, and $\pi(\widetilde{x}, \widetilde{y}) > \pi(\widetilde{y}, \widetilde{x})$.

To check that possibilities (i)–(iii) do not give rise to any incomplete information blocking pairs, simply note that all label-$u\tau$ and label-τ agents obtain their highest possible utility; label-$u\theta$ and label-θ agents do not care about the type of their opponents and thus cannot further improve on their utilities through deviations from NE^{lb}_θ. For all these Bayes-Nash stable outcomes, we have

$$G_\tau(p, \mu, S) > \frac{1}{2}\pi(\widetilde{x}, \widetilde{y}) + \frac{1}{2}(\widetilde{y}, \widetilde{x}) > G_\theta(p, \mu, S).$$

Hence, we have found a preference type τ that is evolutionarily stable against θ. This means θ is evolutionarily unstable. □

O.4 Examples

O.4.1 Assumption on α is Not Redundant in Proposition 3(i)

Consider a material game where each player has three pure strategies. The material payoffs are given as follows

	A	B	C
A	$0,0$	$3,5$	$2,8$
B	$5,3$	$0,0$	$0,0$
C	$8,2$	$0,0$	$0,0$

The strategy pair (A, C) is the unique efficient Nash equilibrium in the material game. It is also a loser-best Nash equilibrium in the material game, i.e. $(A, C) \in NE_{\pi}^{lb}$. Let θ denote the α-homophilic selfish type with $\alpha \leqslant 1$. Let τ be a mutant type that only derives utility from playing (A, C) with its own kind and from playing B against the incumbent type. For example,

$$u_{\tau}(x, y, t) = \begin{cases} 1 & \text{if } (x, y) \in \{(A, C), (C, A)\} \text{ and } t = \tau, \\ 1 & \text{if } x = B \text{ and } t = \theta, \\ 0 & \text{otherwise.} \end{cases}$$

For any $\varepsilon \in (0, 1)$, at state $(\theta, \tau, \varepsilon)$, take a Nash stable outcome (μ, S). We have two possibilities. (i) If $\mu_{\theta,\theta} = \mu_{\tau,\tau} = 1$, then $S = \{(A, C)_{\theta,\theta}, (A, C)_{\tau,\tau}\}$. In this case, $G_{\theta}(\mu, S) = G_{\tau}(\mu, S)$. (ii) If $\mu_{\theta,\tau} > 0$, we must have $(A, B)_{\theta,\tau} \in S$ due to the construction of $u_{\tau}(x, y, t)$. Note that this outcome is Nash stable because type-τ agents already obtain their highest possible utility, and two type-θ agents do not want to deviate since $2 + \alpha \leqslant 3$. In this case, $G_{\theta}(\mu, S) < 5 = G_{\tau}(\mu, S)$. Therefore, when $\alpha \leqslant 1$, the α-homophilic selfish type is evolutionarily unstable.

This example shows that for selfishness to be stable, sufficiently strong homophily is required. Otherwise, when homophilic selfish incumbents are matched with the mutants, they may not have strong enough incentives to rematch and escape a disadvantageous outcome.

O.4.2 (IIIb) Does Not Imply (IIIa) in Definition 12

We now provide an example to show that case (IIIb) in Definition 12 is not redundant. Consider the following material game, and let θ be the parochial efficient type.

	A	B
A	$3,3$	$0,0$
B	$0,0$	$0,0$

Suppose τ is a type that has different preferences against different opponents. The utility function $u_\tau(x, y, t)$ is given by the two matrices in the following tables:

against τ	A	B
A	$0, 0$	$-6, 6$
B	$6, -6$	$1, 1$

against θ	A	B
A	$2, 2$	$0, 0$
B	$0, 0$	$1, 1$

Now consider a population state $(\theta, \tau, \varepsilon = \frac{1}{2})$ and an outcome (p, μ, S) that satisfies: $p_u = 1$, $\mu_{u,u} = 1$, and $S = \{(B, B)_{u,u}\}$. That is, all agents have unobservable types, and they are matched to play the strategy pair (B, B). In the status quo, label-$u\theta$ agents obtain 0 and label-$u\tau$ agents obtain 1. Internal stability is obviously satisfied.

We first argue that there exists *no* incomplete information blocking pair with conditional incentives (i.e. in the sense of case (IIIa) in Definition 12). For either label-$u\theta$ or label-$u\tau$ agents, they are willing to deviate only if they face label-$u\theta$ agents with positive probability. Thus, we have three scenarios to consider:

- $D = D' = \{\theta\}$ and $\boldsymbol{x}(\theta) = \boldsymbol{y}(\theta) = A$. Then label-$u\tau$ agents want to join either side and play A and get 2 which is higher than 1 in the status quo.
- $D = \{\theta, \tau\}$, $D' = \{\theta\}$, and $\boldsymbol{x}(\theta) = \boldsymbol{y}(\theta) = A$. We have $\boldsymbol{x}(\tau) = A$ as it must be a best response. But then label-$u\tau$ agents want to join D' and play B because $\frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 6 > \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 2$.
- $D = D' = \{\theta, \tau\}$ and $\boldsymbol{x}(\theta) = \boldsymbol{y}(\theta) = A$. It is easy to check that label-$u\tau$ agents have a dominant strategy B, which means $\boldsymbol{x}(\tau) = \boldsymbol{y}(\tau) = B$. Then label-$u\tau$ agents receive only $\frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 0 = \frac{1}{2}$ which is less than 1 in the status quo.

Case (IIIb) in Definition 12, i.e. blocking with strong incentives, now has a bite: label-$u\theta$ agents strictly benefit from coordinating on the strategy pair (A, A) *regardless of* whether label-$u\tau$ agents will join and what strategies they will play. Therefore, an incomplete information blocking pair with strong incentives does not imply one with conditional incentives. The challenge with conditional incentives arises from the fact that, despite the parochial efficient agents having a significant motivation to sort themselves out, it is not possible to formulate deviation plans that are compatible with label-$u\tau$ agents' deviating incentives.

O.4.3 Parochial Selfish Type is Not Neutrally Stable

In this section, we construct a material game where the parochial selfish type is not neutrally stable even though all strategy pairs in NE^{lb}_{π} are efficient. Consider a material game where each player has three strategies. The material payoffs are as follows:

	A	B	C
A	$0, 0$	$8, 10$	$7, 10$
B	$10, 8$	$0, 0$	$0, 0$
C	$10, 7$	$0, 0$	$0, 0$

First observe that for this game, $NE_\pi^{lb} = \{(A, B), (B, A)\}$, and both strategy pairs in NE_π^{lb} are efficient. However, there are other Nash equilibria among selfish agents that are inefficient, e.g. (C, A). Now write θ for the parochial selfish type. Consider a type τ that is an "anti-parochial" efficient type who likes to play the game with θ:

$$u_\tau(x, y, t) = \begin{cases} \pi(x, y) + \pi(y, x) & \text{if } t = \tau, \\ \pi(x, y) + \pi(y, x) + 1 & \text{if } t = \theta. \end{cases}$$

Now consider a population state $(\theta, \tau, \varepsilon = \frac{1}{2})$ and an outcome (p, μ, S) as follows. The matching profile (p, μ) satisfies: (i) $p_\theta = p_u = \frac{5}{18}$ and $p_\tau = \frac{4}{9}$; (ii) $\mu_{\theta,\theta} = \mu_{u,u} = \mu_{\tau,u} = \mu_{\theta,\tau} = 0$ and $\mu_{\theta,u} = \mu_{\tau,\tau} = 1$. That is, all label-$\theta$ agents are matched with label-u agents, and label-τ agents are matched among themselves. Note that $q_{u\theta} = \frac{4}{5}$. The strategy profile is $S = \{(C, A)_{\theta,u}, (B, A)_{\tau,\tau}\}$. This outcome is depicted in Figure 5.

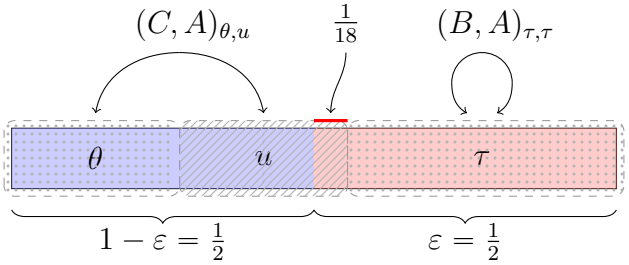

Figure 5: The matching profile (p, μ).

It is easy to verify that S is a Bayes-Nash equilibrium profile. We now argue that there does not exist an incomplete information blocking pair under (p, μ, S). First, a label-τ agent can never attract label-θ or label-$u\theta$ agent to rematch, and she already obtains the highest possible utility when matched with her own kind, so she never participates in a blocking pair. Next, consider the following three cases:

- The best proposal two label-θ agents can make to each other is to coordinate on the loser-best outcome (A, B) (or (B, A)). However, this inevitably creates a "loser" in each pairwise deviation. The loser obtains a utility of 8, which is equal to what label-θ agents can derive in the status quo, $10 \times \frac{4}{5} = 8$. Thus, the blocking pair is not viable.

- Now consider a label-θ agent and a label-$u\theta$ agent. The label-θ agent never benefits

from putting positive probability on strategy A regardless of what the opponent will play, so she could only promise to play B, C, or a mixture. If she promises to play a pure strategy C, a label-$u\theta$ agent will not participate in the deviation. However, if she puts positive probability on strategy B, a label-$u\tau$ would find it profitable to join the pairwise deviation and play the best response A, so the expected utility of the label-θ agent is the same as in the status quo. Therefore, no such incomplete information blocking pair exists.

- Consider two label-u agents contemplating a pairwise deviation. Since $u_\tau(A, C, \theta) = 17 + 1 = 18$, label-$u\tau$ agents participate only if they face label-$u\theta$ opponents with a positive probability. Therefore, we have four scenarios to consider according to cases (IIIa) and (IIIb) in Definition 12:

  First, conditional incentives with $D = D' = \{\theta\}$. By the utility functions of type θ, $(\boldsymbol{x}(\theta), \boldsymbol{y}(\theta))$ must be a Nash equilibrium of the material game. The best symmetric mixed strategy Nash equilibrium they can coordinate on is $(\frac{4}{9}A + \frac{5}{9}B, \frac{4}{9}A + \frac{5}{9}B)$, but they can only derive a utility of $\frac{40}{9} < 7$, even if label-$u\tau$ agents do not participate. Now consider all asymmetric Nash equilibria at once $(A, \alpha B + (1-\alpha)C)$ and suppose $\boldsymbol{x}(\theta) = A$. Then the incentive compatibility of D' is not satisfied because label-$u\tau$ agents also want to participate by playing B and obtain $19 > 18$.

  Second, conditional incentives with $D = \{\theta, \tau\}$ and $D' = \{\theta\}$. If $\boldsymbol{y}(\theta) = A$, the highest utility $\theta \in D'$ can obtain is $8 \cdot \frac{4}{5} < 7$, which is worse than in the status quo. If $\boldsymbol{y}(\theta) = \alpha B + (1-\alpha)C$, then $\boldsymbol{x}(\theta) = \boldsymbol{x}(\tau) = A$, which means τ also wants to join D' because $\frac{4}{5} \cdot 19 + \frac{1}{5} \cdot 18 > 18$, a contradiction.

  Third, conditional incentives with $D = D' = \{\theta, \tau\}$. Then we must have either $\boldsymbol{x}(\theta) = \boldsymbol{x}(\tau) = A$ or $\boldsymbol{y}(\theta) = \boldsymbol{y}(\tau) = A$. Suppose $\boldsymbol{x}(\theta) = \boldsymbol{x}(\tau) = A$ without loss, then the utility of $\theta \in D$ is at most $\frac{4}{5} \cdot 8 < 7$, which means it does not benefit from the deviation.

  Finally, strong incentives for two label-$u\theta$ agents. Similar to the first scenario, we only need to consider asymmetric Nash equilibria $(A, \alpha B + (1-\alpha)C)$ in a deviation. The label-$u\theta$ agents who will play A obtain at most $\frac{4}{5} \cdot 8 < 7$ if label-$u\tau$ opponents join and play B, so they are reluctant to carry out the deviation.

We conclude that (p, μ, S) is a Bayes-Nash stable outcome. The average material payoffs of the two types are computed as

$$G_\theta(p, \mu, S) = \frac{5}{9} \cdot 10 + \frac{4}{9} \cdot 7 = \frac{78}{9},$$

$$G_\tau(p,\mu,S)=\frac{1}{9}\cdot 7+\frac{8}{9}\cdot 9=\frac{79}{9}.$$

Therefore, $G_\theta(p,\mu,S)<G_\tau(p,\mu,S)$, and thus the parochial selfish type is not neutrally stable in this example.

O.4.4 Heterophilic Matching in Polymorphic Population is Not Robust

Suppose the material payoffs of the underlying game is given as follows:

	A	B
A	$0,0$	$5,1$
B	$1,5$	$0,0$

Consider a population distribution ν with $\nu(\tau)=\nu(\tau')=0.5$. Suppose $u_\tau(x,y)=\pi(x,y)+\pi(y,x)+\mathbb{1}_{\{t\neq\tau\}}$ and $u_{\tau'}(x,y)=\pi(x,y)+\pi(y,x)+2\mathbb{1}_{\{t\neq\tau'\}}$. So both types have a preference for efficiency and they are heterophilic in the sense that they derive an extra utility from interacting with agents that are different from them. Consider an outcome $\mu_{\tau,\tau'}=1$ (the matching is perfectly heterophilic), and a strategy profile $S=\{(A,B)_{\tau,\tau'}\}$. The matching profile $(\mu(\nu),S)$ is uniquely Nash stable. However, The two types do not earn the same material payoffs because the efficient strategy pair (A,B) is asymmetric. Hence, the condition on the material game for ν to be balanced (the first criterion for evolutionary stability) is not satisfied.

Suppose we relax the assumption that the strategies played in an outcome are pairwise-homogeneous. That is, the strategy pair played by two matched agents only depends on the types of those agents. Instead, we allow exactly half of the τ-τ' pairs play (A,B) and the rest play (B,A). In this case, both types get an average material payoff of 3, so the population is balanced. However, we will show that the population is susceptible to mutations.

Consider a mutant type τ'' whose utility function exhibits plasticity and is given as follows:

		type τ''	
		A	B
type τ''	A	$0,-$	$0,-$
	B	$1,-$	$1,-$

	non-type τ''	
	A	B
A	$1,-$	$1,-$
B	$0,-$	$0,-$

A τ''-type agent has a dominant strategy of playing B against another τ''-type agent (so the τ''-type exhibits same-type inefficiency because they play (B,B) and earn a material payoff of 0 when matched among themselves), and has a dominant strategy of playing A against any agent that is not type τ''. Consider a post-entry population state $\widetilde{\nu}=(1-\varepsilon)\nu+\varepsilon\tau''$, a matching profile $\mu(\widetilde{\nu})$ in which all the τ''-type agents are matched with the τ'-type agents

(in total ε pairs), the rest of the τ'-type agents are matched with the τ-type agents (in total $(1-\varepsilon)/2-\varepsilon$ pairs) and the leftover τ-type agents are matched among themselves (in total $\varepsilon/2$ pairs); and a strategy profile S in which all the τ-τ pairs play (A,B), half of the $\tau-\tau'$ pairs play (A,B), while the rest τ-τ' pairs play (B,A), and all the $\tau'-\tau''$ pairs play (B,A). The outcome $(\mu(\widetilde{\nu}),S)$ is Nash stable because 1) all the τ-type and τ'-type agents are playing an efficient strategy profile with their opponents and the τ''-type agents are playing their dominant strategies; 2) all the τ'-type agents get a utility of 8 and all the τ''-type agents get a utility of 1, so they are not willing to form a blocking pair with those τ-type agents in the $\tau-\tau$ matches.

Given this Nash stable outcome for the post-entry population state, $G_{\tau'}(\mu(\widetilde{\nu}),S) = 3*(1-3\varepsilon)/(1-\varepsilon)+1*2\varepsilon/(1-\varepsilon)<3=G_{\tau''}(\mu(\widetilde{\nu},S)$. Hence, ν is not evolutionarily stable against τ'', and it is evolutionarily unstable. The rationale for ν being evolutionarily unstable is that in the above described Nash stable outcome, some τ'-type agents are matched with τ''-type agent whose dominant strategy is A, the more advantageous strategy in the efficient strategy pair (A,B). Hence, these τ'-type agents would best respond by choosing B because of their preference for efficiency, which gives them a low material payoff of 1.

Note that if all efficient strategy pairs in the underlying material game are symmetric, then the population distribution ν that contains equal proportions of the two heterophilic preference types τ and τ' is evolutionarily stable against any mutant that exhibits same-type inefficiency because the efficiency strategy pairs being symmetric prevents the incumbents from being taken advantage of by the mutants, and there always exists a Nash stable outcome for the post-entry population such that every τ is matched with another τ' to play efficiently, while the mutant type agents are matched among themselves playing inefficiently.

In sum, it is difficult for a polymorphic population distribution that consists of heterophilic preference types and features heterophilic matching to satisfy the balance condition. Even it does, it can be invaded. The only scenario in which it can be evolutionarily stable against any mutant type that exhibits same-type efficiency requires all efficient strategy pair to be symmetric in the underlying material game.

Of course, a polymorphic population consisting of various homophilic and parochial efficient type agents can be evolutionarily stable. Nevertheless, in such a population, matching is assortative instead of heterophilic. In sum, by allowing polymorphism, heterophilic preferences and heterophilic matching are not as robust as homophilic preferences and assortative matching under evolutionary selection pressure.