UNIVERSITY OF LEICESTER

Department of Economics

# A Simple Model of Homophily in Social Networks

**Sergio Currarini,** University of Leicester
**Jesse Matheson,** University of Leicester
**Fernando Vega Redondo,** Bocconi University

Working Paper No. 16/05

# A Simple Model of Homophily in Social Networks

Sergio Currarini* Jesse Matheson† Fernando Vega Redondo‡

March 20, 2016

**Abstract**

Biases in meeting opportunities have been recently shown to play a key role for the emergence of *homophily* in social networks (see Currarini, Jackson and Pin 2009). The aim of this paper is to provide a *simple* microfoundation of these biases in a model where the size and type-composition of the meeting pools are shaped by agents' socialization decisions. In particular, agents either inbreed (direct search only to similar types) or outbreed (direct search to population at large). When outbreeding is costly, this is shown to induce stark equilibrium behavior of a threshold type: agents "inbreed" (i.e. mostly meet their own type) if, and only if, their group is above certain size. We show that this threshold equilibrium generates patterns of in-group and cross-group ties that are consistent with empirical evidence of homophily in two paradigmatic instances: high school friendships and interethnic marriages.

*Keywords: Homophily, social networks, segregation.*

*JEL Classification: D7, D71, D85, Z13.*

*Department of Economics, University of Leicester and Universita' Ca' Foscari di Venezia. Email: sc526@le.ac.uk.
This author wishes to acknowledge the support of the Ministry of Education and Science of the Russian Federation, grant No. 14.U04.31.0002, administered through the NES CSDSI.
†Department of Economics, University of Leicester. Email: jm464@le.ac.uk.

‡Bocconi University and IGIER. Email: fernando.vega@unibocconi.it.

# 1 Introduction

A pervasive feature of social and economic networks is that contacts tend to be more frequent among similar agents than among dissimilar ones. This pattern, usually referred to as "homophily", applies to many types of social interaction, and along many dimensions of similarity.[1] The presence of homophily has important implications on how information flows along the social network (see, for example, Golub and Jackson (2011)) and, more generally, on how agents' characteristics impinge on social behavior. It is therefore important to understand the generative process of homophilous social networks, and how agents' preferences and their meeting opportunities concur in determining the observed mix of social ties.

The empirical evidence of many social networks shows that homophily is often in excess of the "baseline" level that would be expected under a uniform random assortment that reflected groups' population shares, and that inbreeding (within-group interaction) occurs both in small and large groups. In a seminal contribution, Currarini, Jackson and Pin (2009)—CJP henceforth—investigate the extent of such biases in the context of American high school friendships. They show that preferences which are biased in favor of same-type friendships help explain why member of large ethnic groups enjoy more popularity—a higher number of friends—than members of small ethnic groups. However, they find that the observed patterns of inbreeding homophily *cannot* be explained by a process in which agents meet purely at random. They conclude, therefore, that some kind of "meeting bias" must be at work.[2]

In this paper we study a micro-founded model of search which endogenously generates a meeting bias, and whose equilibrium predictions are consistent with the observed non-linear relationship between homophily and population shares common to both U.S. high school friendship nominations and U.S. marriages. A characterizing, and novel, feature of our model is the role of absolute group size in shaping agents incentives to either direct their search towards in-groups only, or to open up to interactions with out-groups as well. This marks a stark difference between our approach and the approach based on the role of population shares, central to all previous studies of homophily in economics and to Blau (1977)'s structural approach. We will discuss this difference in some detail in Sections 4 and 5. We test our model's predictions on the role of group size using micro data reflecting two different matching scenarios: friendship nominations and marriages. The empirical results support our model. We now present in some detail the model, and then discuss our contribution with respect to recent works on the subject.

The essential features of our theoretical framework can be outlined as follows. Agents derive

[1] For an account of the pervasiveness of homophily, see the seminal work of Lazarsfeld and Merton (1954) and, more recently, Marsden (1987, 1988), Moody (2001), or the survey by McPearson, Smith-Lovin and Cook (2001).

[2] In particular, random meetings are shown to be inconsistent with the nonlinear relation between an index of homophily first proposed by Coleman (1958) and groups' population shares.

positive utility from the number of *distinct* ties they enjoy. Ties are formed from a fixed number of meeting draws obtained from an endogenously chosen meeting pool. Agents affect the composition of their meeting pool by choosing to either *inbreed* or *outbreed*. Inbreeding refers to the decision to restrict search to one's own group only; outbreeding refers to the decision to extend search to the whole population. The decision to inbreed or outbreed involves weighing conflicting incentives. Outbreeding is costly; we believe this reflects cultural, geographical, or linguistic barriers to accessing other types. Inbreeding limits the size of the search pool and, therefore, the efficacy of search by affecting the probability of novel draws.[3]

An agent's breeding decision depends crucially on the size of her group, to the extent that this affects the probability of redundancies in search. Specifically, there exists a threshold group size above which the agent will inbreed and below which the agent will outbreed. We highlight two paradigmatic scenarios that embody polar assumptions on how agents connect. The first scenario involves a meeting mechanism where links and payoff flows are one-sided. This represents, for example, web-based social networks (such as Twitter) where links are directed and information flows in one direction. This scenario also captures, to some extent at least, the friendship nomination process on which the National Longitudinal Study of Adolescent Health is based. The second scenario involves a meeting mechanism where both connections and payoff flows are two-sided; links require some form of bilateral agreement or coordination. Marriages (mutual consent being required) are a natural example of this scenario.

We show in Section 4 that the threshold equilibrium, together with some small random noise in meetings, predicts a qualitative pattern of the Coleman index which is consistent with the hump shaped pattern found in CJP for friendship nomination and also arising for U.S. marriages (see Figure 1 in the present paper). In addition, we show that focusing on the role of absolute group size helps explain differences in the aggregate homophily patterns of small vs. large schools that were identified in Currarini, Jackson and Pin (2010) but could not be explained in their framework (in large schools, the degree of homophily is uniformly higher). Using *microlevel data* on friendship nominations and marriages, in Section 5 we test other novel predictions of the model. One such prediction is that, conditional on relative population share and both in the one-sided and two-sided scenarios, inbreeding is more likely to occur in groups that are large in absolute size than in smaller groups. Another interesting theoretical prediction for which we find empirical support pertains, specifically, to the matching performance of small groups. It concerns the following contrast between one- and two-sided contexts. If matching is one-sided, the matches of any small group will have all other groups (large or small) represented according to their population shares. Instead, if matching is two-sided, the prediction is that outbreeders will meet each other with frequencies that reflect

[3]So, while in CJP the focus is on agents' decision of how intensively (i.e. for how long) to search for social ties, while the meeting probabilities are fixed exogenously (and hence outside of agents' influence), in our case all agents search with the same intensity, but are able to direct their search and thus affect their meeting probabilities.

their population shares *within the pool of outbreeders*. As a result, we have that outbreeding groups will be over-represented in the matches of other outbreeders relative to their population shares. We believe that all these findings provide strong empirical support to our model.

The general idea that homophily patterns may stem from selection and assortative matching is present in many theoretical constructs and has been extensively tested empirically since Kandel's (1978) work on adolescent friendships. In Tiebout's "voting-by-feet" model, agents selectively structure their social interactions by forming homogeneous clubs along the preference dimension. The anticipation of future interaction is also at the heart of Baccara and Yariv (2013), where homophilous peer groups form in connected intervals along the preferences dimension. Selection may also result from information and opinion seeking, as in Suen (2010) mutual admiration clubs, where similar agents communicate in a sort of self confirming updating of information. Selection of agents with similar preferences may also stem from the desire to avoid strategic manipulation of information, as in Galeotti et al. (2013) model of cheap talk in networks. There is also a similarity between the main feature of our process (the difference in the cost of linking with in-group versus out-group agents) and the approach taken in Jackson and Roger (2005)'s islands model of network formation. However, while in that paper a key role is played by indirect benefits and the focus was on the emergence of small world architectures, here the focus is on the effect of different group sizes on homophily patterns in the absence of indirect benefit from connections.

The importance of absolute group size has been put forward before in the attempt to explain why groups with small relative population shares may end up displaying high homophily. Moody (2001)'s discussion of U.S. high school friendships contains in fact the main elements of our analysis: "...(s)ince people have a finite capacity for relationships (van der Poel 1993; Zeggelink 1993), in a school where there are many minority students, minority students may be able to find their desired number of friends within the minority friendship pool." In Section 3 we show that our model predicts that minority groups members, sampling a limited number of friends, will stick to their own pool as long as its absolute size does not impose too large inefficiencies on search. This happens when minorities are large in absolute terms; that is, in schools with large total populations (see Fig. 2 in the present paper). Similar insights are contained in the discussion of inter-religious marriages in Fisher(1992), where it is reported how "...resident of small towns risk falling away from their religious roots, presumably because co-religionists are less likely to be available, while resident of larger cities are more likely to be enveloped in a religious sub-culture". The empirical finding that inbreeding is often used by small minorities to counteract the averse effect of relative population share on in-group ties (see McPherson et al., 2001) seems consistent with the results of our regressions on micro-data, where we find that the marginal effect of releasing the constraint of absolute size on the tendency to inbreed are larger the smaller is the relative population share of the group, signalling an overall stronger tendency to inbreed of small minorities when this comes at little cost in terms of search efficiency.

The remainder of the paper is organized as follows. In Section 2 we describe the model, including the strategies and payoffs defining the underlying meeting game. In Section 3 we characterize the equilibrium behaviour of agents, as a function of the size of their respective groups. In Section 4 we discuss the aggregate empirical patterns of homophily for friendships and marriages, and derive the main results that link our theory to the observed empirical patterns. In Section 5 we use micro-data to test the novel predictions of the theory. Finally, in Section 6 we conclude the main body of paper with a summary. For the sake of exposition, all proofs are included in the Appendix 1 and data summaries are included in Appendix 2.

# 2 The model

## 2.1 Framework

Consider a set $N \subset \mathbb{N}$ of $n$ agents. The set $N$ is partitioned into $q$ groups, defined by a specific common trait (ethnic, linguistic, religious, etc.), which we call "type". Groups are indexed by $l$, and we denote by $n_l$ the (absolute) size of group $l = 1, 2, ..., q$. Let us also consider a network defined on the set $N$, where we use the terms "match" as synonymous of "link". Now assume that each agent $i \in N$ devotes a fixed amount of time to meet other agents in $N$. In this lapse of time he obtains $\eta > 1$ random draws with replacement. Out of these draws, let $\nu\,(\leq \eta)$ denote the number of distinct agents he meets. In the end, not all of the distinct agents $i$ meets turn out to be suitable partners. We assume that this happens, in a stochastically independent manner for each of them, with probability $p$ $(0 < p < 1)$.

In this context, the sole decision every agent must take is how to allocate time between meeting agents of her own group and agents of the whole population (including her group). The first type of activity is referred to as "inbreeding", and the second as "outbreeding". We assume that, in order for outbreeding to be feasible, the agents must incur a fixed cost $c$. This cost can be interpreted as reflecting some form of investment required to interact with people of different groups (e.g., travelling, learning a language, or changing one's habits).

The inbreeding/outbreeding decisions taken by all agents constitute their strategies in the matching game. They determine the meeting pool each of them accesses, which in turn shapes the probability distribution over the number of *distinct* partners they face, and thus their expected payoffs.

## 2.2 Shaping the meeting pool

In general, the meeting pool faced by any given agent is a consequence of her own breeding decision, as well as that of all others. Denote by $I$ and $O$ the inbreeding and outbreeding decisions, respectively. Then, in principle, the meeting pool of each agent is a set-valued function $\Theta_i(\mathbf{s})$ of the profile $\mathbf{s} \equiv (s_i)_{i \in N} \in \{I, O\}^N$ specifying the breeding decisions of all agents. The cardinality of $\Theta_i(\mathbf{s})$, measuring the size of the meeting pool, is denoted by $\theta_i(\mathbf{s})$. Given any profile $\mathbf{s}$, the random variable $\tilde{\nu}(\eta, \theta_i(\mathbf{s}))$ specifies the number of distinct partners obtained from $\eta$ uniform and independent draws with replacement, when the *size* of the pool is $\theta_i(\mathbf{s})$.

As advanced, we shall distinguish two different scenarios (one-sided and two-sided) concerning how the meeting pool of an agent is shaped by the strategy profile $\mathbf{s}$. Each scenario is captured by specific forms for the functions $\theta_i(\cdot)$ and $\tilde{\nu}(\cdot)$ that shape, respectively, the size of the meeting pool and the meeting opportunities.

**(a) One-sided Scenario**

The simplest case is given by a meeting scenario that is *one-sided*, in the sense that the conditions enjoyed by any given agent exclusively depend on her own choices and her own meeting draws. It can be used to model situations in which the formation of a tie is fully determined by the initiator of the tie, while the receiving agent has no control over it. As suggested before, this includes those empirical setups where friendship is recorded through individual (independent) nominations and two agents are identified as friends when at least one of them lists the other. One-sidedness is also a feature displayed by those contexts where connections are established in order to acquire information in a strictly unilateral endeavour. This happens, for example, in internet browsing (where inbreeding may mean that an agent only connects to blogs of similar political orientation or nationality) or in certain social networks (such as those supported by Twitter) where virtual friends/followers cannot be refused.

To formalize matters, denote by $l(i)$ the index of the group to which agent $i$ belongs, and let $\overrightarrow{\theta_i}(\mathbf{s})$ denote the size of the meeting pool accessed in this context by that agent, given the strategy profile $\mathbf{s}$. Then we posit:[4]

$$\overrightarrow{\theta_i}(\mathbf{s}) = \begin{cases} n_{l(i)} & \text{if } s_i = I \\ n & \text{if } s_i = O. \end{cases} \tag{1}$$

where recall that $n_{l(i)}$ stands for the cardinality of group $l(i)$.

In line with the postulated one-sidedness of the meeting mechanism in this case, the payoff

[4]For notational simplicity, agent $i$ is included in the pool, even though she cannot obviously meet herself. The same simplification is applied below to the two-sided scenario.

flows will be assumed to be one-way as well.[5] By this it is meant that payoffs accrue only to the agent who actively finds a suitable partner but *not* in the opposite direction. Thus, *ex ante*, the (uncertain) distribution of payoffs is governed by the random variables $\tilde{\nu}_{\rightarrow}(\eta, \theta)$ that give the number of distinct draws out of $\eta$ tries when the pool size is $\theta$.

**(b) Two-sided Scenario**

In contrast to the previous case, a *two-sided* context is one where the formation of a tie requires the consent of both parties involved, in the sense that both of them have to choose to be part of the same meeting pool. The case of marriages falls clearly into this class of situations. In some cases, the outbreeding choice may be implemented by moving to some fixed location ("downtown") where individuals from different groups meet, or by switching to a common *lingua franca* that is different from the group's native language. Formally, as before, meeting pools are determined by agents' inbreeding/outbreeding decisions. Now, however, outbreeding agents only access (besides those of their own group)[6] the agents of other groups that have themselves chosen to outbreed. This gives rise to an alternative function $\overleftrightarrow{\theta_i}(\mathbf{s})$ specifying the size of the meeting pool. Let $n_l^I(\mathbf{s})$ and by $n_l^O(\mathbf{s})$ denote the number of agents of type $l$ that choose the strategy $I$ and $O$ in $\mathbf{s}$, respectively. Then we have:

$$\overleftrightarrow{\theta_i}(\mathbf{s}) = \begin{cases} n_{l(i)}^I(\mathbf{s}) & \text{if } s_i = I \\ \sum_{l=1}^{q} n_l^O(\mathbf{s}) & \text{if } s_i = O. \end{cases} \tag{2}$$

When matching is two-sided, it it is natural to assume that the payoff flows are two-way, i.e. a link established by two suitable partners generates positive payoffs to both of them. The random variable used to account for this must therefore be different from the one used for the one-sided scenario. The present one, denoted by $\tilde{\nu}_{\leftrightarrow}(\eta, \theta)$, gives the random number of distinct meetings obtained in a pool of $\theta$ agents when: $(a)$ each agent makes $\eta$ independent draws with replacement; $(b)$ a meeting is said to occur between two agents, $i$ and $j$, when either a draw by $i$ selects $j$ or *vice versa*.

In both one-sided and two-sided scenarios, a larger pool obviously brings about richer meeting possibilities – i.e. the range of distinct partners an agent can meet is wider. It is this simple feature

[5] The distinction between one-sided and two-sided link formation and the (conceptually different) contrast between one-way and two-way flows is discussed at length in Bala and Goyal (2000), one of the earliest papers of the network formation literature in economics.

[6] Given that our analysis focuses on symmetric equilibria (see Subsection 2.4), all agents of any given group must belong to the same matching pool. This, however, could be easily generalized since, as the population gets large, the relative size of a group whose size remains bounded becomes infinitesimal. Therefore, the situation faced by outbreeders would be essentially the same whether all individuals of their group are outbreeders or not and hence our main conclusions would still hold in the absence of group-based symmetry.

that introduces the basic tradeoff between the inbreeding and outbreeding decisions that is at the core of our model.[7] Mathematically, such richer possibilities are captured by the fact that the two families of random variables, $\{\tilde{\nu}_{\rightarrow}(\eta,\theta)\}_{\theta\in\mathbb{N}}$ and $\{\tilde{\nu}_{\leftrightarrow}(\eta,\theta)\}_{\theta\in\mathbb{N}}$, can be suitably ranked when parametrized by the size of the meeting pool $\theta$. Indeed, we will show that those random variables are strongly ordered as follows:

- In the one-sided scenario, larger meeting pools yield probability distributions over the number of distinct draws that dominate those of smaller pools in the First-Order Stochastic Dominance sense.
- In the two-sided scenario, larger meeting pools yield a expected number of distinct draws that is higher than for smaller ones.

In general, we can conceive either the one- or the two-sided scenarios as a more appropriate modeling choice depending on the characteristics of the situation. (For example, in Section 5 we suggest that the first approach is more in line with our friendship data while the second is more consistent with our data on marriages.) However, both modeling alternatives induce, under suitable assumptions on preferences,[8] a similarly sharp trade-off between the inbreeding and outbreeding options – cf. Theorems 1 and 2.

### 2.3 Preferences

Now we describe agents' preferences over their meeting outcomes. Denote the number of distinct meetings enjoyed by any given agent $i$ by $\nu_i$. Given that each distinct partner happens to be suitable with probability $p$, the induced number of suitable partners, denoted by $y_i$, is given by the Binomial distribution $\mathbf{Bin}(\nu_i, p)$. We assume that agents evaluate that (uncertain) outcome according to some common von Neumann-Morgenstern (vNM) utility $U : \mathbb{N}\cup\{0\} \rightarrow \mathbb{R}$, where we normalize $U(0) = 0$ and posit that

$$U(y_i+1) \geq U(y_i) \qquad \text{for all } y \in \mathbb{N} \tag{3}$$

$$U(1) > U(0). \tag{4}$$

[7] Other specifications of the matching mechanism that display this trade-off yield conclusions that are qualitatively the same as in our postulated one- and two-sided meeting scenarios. By way of illustration, let us sketch two examples. In the first one, agents enjoy partner variety, which is in line with standard assumptions on preferences made in economic theory. If we then make the natural assumption that such a variety grows in expectation as the pool of alternative partners expands, the desired effect of pool size follows. As a second possibility, suppose that new partners are searched through existing partners (i.e. as friends of friends) in a random social network whose size depends (as in our model) on the breeding decisions taken. Then, the effectiveness of such search depends on network clustering, which in turn is well know to decrease with size in random networks (see e.g. Vega-Redondo (1997)).

[8] Naturally, as we shall see, assumptions on preferences must be stronger in the second case since the corresponding dominance criterion that is used is weaker.

Thus the assumption is that the utility does not fall as the number of suitable partners grows, with a strict improvement only required when passing from a situation with no suitable partner to one with some such partner. In general, therefore, the theoretical framework may accommodate different applications, such as friendships or marriages. For example, in the former case, it would be natural to posit that $U$ is strictly increasing throughout, while in the second case it may be postulated to level at one.[9]

Next, we can define the expected utility $V(\nu_i)$ induced by any given number $\nu_i$ of distinct partners of agent $i$ as follows:

$$V(\nu_i) \equiv \sum_{y_i=0}^{\nu_i} \binom{\nu_i}{y_i} \; p^{y_i} \; (1-p)^{\nu_i - y_i} \; U(y_i). \tag{5}$$

Finally, we take into account the fact that, given the pool size $\theta_i$ faced by an agent $i$, the number $\nu_i$ of her distinct partners is uncertain from an *ex ante* viewpoint. It is determined by the random variable $\tilde{\nu}(\eta, \theta_i)$, as particularized to the scenario under consideration (one- or two-sided). We thus need to integrate (5) with the distribution over the number of distinct partners induced by pool size $\theta_i$. This gives rise to the expected utility $W(\theta_i)$ for a typical agent $i$ is defined as follows:[10]

$$W(\theta_i) \equiv \mathbb{E}_{\tilde{\nu}(\theta_i)} V(\nu_i) = \sum_{\nu_i=0}^{n} P_{\theta_i}(\nu_i) \; V(\nu_i). \tag{6}$$

where $P_{\theta_i}(\cdot)$ denotes the probability distribution associated with the random variable $\tilde{\nu}(\theta_i)$ that specifies the number of distinct meetings in the scenario under consideration.

### 2.4 The breeding game

We are now in a position to define the "breeding game." This requires specifying both the strategy sets and the payoff functions.

First, the strategy space of every player $i$ is simply identified with the set $\{I, O\}$ consisting of the two possible breeding decisions she can take: inbreed and outbreed, respectively.

Second, the payoff of the agent is defined as follows:

$$\pi_i(\mathbf{s}) = \mathbb{E}_{\tilde{\nu}(\theta_i(\mathbf{s}))} V(\nu_i) = \sum_{\nu_i=0}^{n} \left\{ P_{\theta_i(\mathbf{s})}(\nu_i) \left[ \sum_{y_i=0}^{\nu_i} \binom{r_i}{y_i} \; p^{y_i} \; (1-p)^{\nu_i - y_i} \; U(y_i) \right] \right\} - c(s_i)$$

[9] Note that our model assumes that the type composition of an agent's meetings is inessential for utility (which only depends on the total number of meetings). While this is meant to isolate the effect of size on meeting possibilities, homophilous preferences could be considered in the model without loosing the key mechanisms behind our results.

[10] For notational simplicity, we henceforth dispense with the parameter $\eta$, since it will remain fixed throughout our analysis.

where (as explained in Subsection 2.1), $c(s_i) = c > 0$ if $s_i = O$ and $c(s_i) = 0$ if $s_i = I$, and the notation needs to be particularized to the scenario being considered (one- or two sided).

Our equilibrium analysis will focus throughout on profiles $\mathbf{s}$ that are group-symmetric,[11] i.e. where $s_i = s_j$ whenever $l(i) = l(j)$. Within this class, the population behavior can be fully described by the $q$-tuple $\boldsymbol{\gamma} \equiv (\gamma_1, \gamma_2, ..., \gamma_q)$ that specifies the common choice $\gamma_l \in \{I, O\}$ for every agent in each of the groups $l = 1, 2, ..., q$. Then, denoting by $\{\theta_l(\boldsymbol{\gamma})\}_{l=1,2,...,q}$ the induced meeting pools, the payoff of any typical agent $i$ of group $l$ is given by:

$$\pi_l(\boldsymbol{\gamma}) = \mathbb{E}_{\tilde{\nu}(\theta_l(\boldsymbol{\gamma}))} V(\nu_i) = \sum_{\nu_i=0}^{n} \left\{ P_{\theta_l(\boldsymbol{\gamma})}(\nu_i) \left[ \sum_{y_i=0}^{\nu_i} \binom{r_i}{y_i} p^{y_i} (1-p)^{\nu_i - y_i} U(y_i) \right] \right\} - c(\gamma_l).$$

# 3 Equilibrium

Here we characterize the group-symmetric Nash equilibria of the game. The key feature to highlight is that the behavior of a group at equilibrium is fully dependent on whether the group is large or small, relative to a certain threshold defined by the equilibrium. More specifically, we find that all groups whose size is smaller than such a threshold outbreed, while larger groups inbreed. This is the equilibrium pattern arising both in the one- and in the two-sided scenarios, but a significant difference exists between them. In the one-sided case, we show that the equilibrium threshold is unique. Instead, in the two-sided context, there is generally a range of possible thresholds that can be supported at equilibrium, a reflection of the unavoidable "coordination problem" that agents face in this case.

We start by characterizing the equilibrium for the one-sided scenario.

THEOREM **1 (Threshold Equilibrium – one-sided scenario)** *Consider the one-sided scenario and assume that the outbreeding cost satisfies $c < V(\eta)$. Then, there exists some $\hat{n}$ and a specific (finite) $\tau^* \geq 2$ such that if $n \geq \hat{n}$, the strategy profile $\gamma^* = (\gamma_l^*)_{l=1}^q$ satisfying:*

$$\gamma_l^* = I \Leftrightarrow n_l \geq \tau^* \qquad (l = 1, ..., q) \tag{7}$$

*defines the unique group-symmetric Nash equilibrium of the breeding game.*

The previous result builds upon the fact that the smaller is a group, the higher the risk faced by its members that, if they restrict to their own kind alone, their meetings may be wasteful (i.e.

[11] The restriction to symmetric equilibria is natural in our case, where all agents in every group are homogeneous (i.e. *ex ante* identical). Homogeneity and the nature of the strategic situation imply that asymmetric equilibria are either non-generic (in the one-sided scenario) or unstable (in the two-sided scenario).

redundant). This then leads to the conclusion that optimal behavior should be of threshold type in group size. Indeed, a key step in the proof of Theorem 1 is showing that the higher redundancy risk associated to smaller inbreeding groups is suitably captured by the strong criterion of First-Order Stochastic Dominance (FOSD). More precisely, we shall prove the following auxiliary lemma.

LEMMA **1** *For any given $\theta$, $\theta'$, if $\theta \geq \theta'$ the random variable $\tilde{\nu}_{\rightarrow}(\theta)$ dominates $\tilde{\nu}_{\rightarrow}(\theta')$ in the FOSD sense.*

One may worry, however, that the threshold $\tau^*$ established by Theorem 1 may be so low that the maximum group size leading to outbreeding is very small. In general, of course, this must depend on the cost $c$ of outbreeding. But it is straightforward to see that if the outbreeding cost $c$ is low enough, the equilibrium threshold can be made arbitrarily large. For completeness, we state this conclusion in the following corollary:

COROLLARY **1** *Under the assumptions made in Theorem 1, for any $\tau_0$ there is some $\hat{n}$ and $\bar{c} > 0$ such that if $n \geq \hat{n}$ and $c < \bar{c}$ then the equilibrium threshold $\tau^* \geq \tau_0$.*

An idea analogous to that underlying the one-sided scenario applies the two-sided case, but with an important caveat already advanced: the benefit of outbreeding now depends on the *endogenous* size of the outbreeders' pool. Hence, in contrast with the one-sided scenario, the game now displays equilibrium multiplicity. To see this, consider, for example, the situation where no group outbreeds, independently of its size. Such a situation obviously defines an equilibrium. For, no matter how small the outbreeding (positive) cost might be, no individual can find it optimal to pay it if the pool of those who outbreed consists only of agents of their own type alone. Despite the possibility of such a "deadlock," the result below establishes that as long as (a) *all* the small groups (as identified for the one-sided scenario) command *in total* a non negligible share of the whole population, and (b) the utility $U$ over the number of distinct meetings is linear, the existence of a positive-threshold equilibrium holds as well in a two-sided scenario.

THEOREM **2 (Threshold Equilibrium – two-sided scenario)** *Consider the two-sided scenario, and assume that the utility function $U$ is linear and the outbreeding cost satisfies $c < V(\eta)$. Then, every group-symmetric equilibrium $\tilde{\gamma} = (\gamma_l)_{l=1}^{q}$ is of the threshold type, i.e. there exists a $\tilde{\tau}$ such that*

$$\tilde{\gamma}_l = I \Leftrightarrow n_l \geq \tilde{\tau} \qquad (l = 1, ..., q).$$

*Moreover, given any $\alpha > 0$ and the threshold $\tau^*$ given in (7), there exists some $\hat{n}$ such that if $n \geq \hat{n}$ and $\sum_{l:n_l<\tau*} n_l > \alpha n$, a threshold equilibrium exists with $\tilde{\tau} = \tau^*$.*

As for the one-sided scenario, the present result builds upon the fact that, under the (stronger) assumptions it contemplates, increasing group size introduces a well-defined ranking on the (stochastic) prospects faced by the corresponding agents. We shall use, specifically, the following lemma.

LEMMA **2** *For any given $\theta$, $\theta'$, if $\theta \geq \theta'$ the expected values of the corresponding random variables, $\tilde{\nu}_{\leftrightarrow}(\theta)$ and $\tilde{\nu}_{\leftrightarrow}(\theta')$, satisfy $\mathbb{E}\left[\tilde{\nu}_{\leftrightarrow}(\theta)\right] \geq \mathbb{E}\left[\tilde{\nu}_{\leftrightarrow}(\theta')\right]$.*

In contrast with Lemma 1, the previous Lemma 2 weakens the criterion used to rank the meeting distributions.[12] This weakening is motivated by the fact that, when matching is two-sided, the size of the pool affects the (random) number of distinct meetings in two different ways. First, a larger pool renders the search for new partners more effective by reducing the likelihood of wasteful redundancies. This is just as in the one-sided model. There is, however, a second dimension of two-sided matching that works in the opposite direction: as the matching pool grows, the probability of being found by any other given agent decreases. These two conflicting considerations make it difficult to analyse the overall effect of a varying pool size on expected utility unless preferences are suitably restricted. A natural such restriction is to posit that agents' risk aversion is limited – or, as contemplated in Theorem 2, that agents are risk neutral and their utility function $U$ linear.[13] Risk neutrality is admittedly a strong assumption, and may fail to capture situations where agents are willing to trade off a larger expected number of matches for a more "stable" distribution. With strongly risk averse agents, for instance, we could not rule out the possibility that members of very small groups prefer to inbreed in order to avoid the risk of very few matches due to the difficulty of being found in larger pools, while members of larger groups may decide to outbreed. We refer, however, to our discussion in footnote 16, suggesting that such possibility should not occur.

To sum up, our analysis of the one- and two-sided scenarios show qualitatively similar threshold behavior. Nevertheless, as emphasized, a key difference between the two cases is that the latter one admits equilibrium multiplicity and, consequently, opens up the possibility of acutely inefficient equilibria embodying miss-coordination. Intuitively, it is quite clear that all equilibria associated to thresholds $\tau < \tau^*$ embody a certain extent of coordination failure. Indeed, that such a failure is indeed a possibility is well illustrated by the fact that, as explained above, there *always* exists the

[12]Note, of course, that when a distribution dominates another one in the FOSD sense, it also yields a higher expected value.

[13]We have investigated the possibility that a ranking based on FOSD, as the one postulated in Lemma 1 for one-sided matching, also applies to the two-sided model. While we could not obtain a general analytical result, we were able to show using numerical analysis that the cumulative distribution of "passive" matches (those through which an agent is found by other agents in the pool) decreases in the pool size $\theta$. Given that the distribution of total matches is the convolution of passive and active matches, using results on the preservation of FOSD in convlolutions (see Lemma 2.1 in Aubrun and Nechita (2009)), we conclude that numerical simulations suggest that the stronger result of Lemma 1 also applies to the two-sided model, and the stronger assumption of risk neutrality could be dispensed of in Theorem 2.

extreme equilibrium with full-inbreeding, induced by a threshold $\tau = 0$. The entailed coordination problem is out the scope of this paper, so we choose to abstract from it by assuming throughout that the highest threshold $\hat{\tau}$ consistent with equilibrium is played in the two-way scenario. As stated in Theorem 2, if the population is large enough, we have $\hat{\tau} \geq \tau^*$.

# 4 Equilibrium implications for homophily

Empirical evidence on the patterns of homophily has traditionally focused on its variation across groups making up for different shares of a total population, or *relative* group size. This makes sense as populations shares are a benchmark measure of expected homophily when social contacts are made at random. Departure of homophily from population share signals some form of bias with respect to uniform assortment. So, although the novelty of the present model is on the role that *absolute* group size plays in determining incentives to outbreed and, therefore, the individual and aggregate patterns of homophily, in this section we focus on some documented stylized facts regarding the relationship between relative group size and homophily. These facts have been interpreted in CJP's (2009) analysis of friendships as evidence of substantial meeting biases at work, and motivate the present exercise to micro-found such biases. We will show that, at the aggregate level, the mechanism proposed in this paper is consistent with the observed pattern of homophily both in friendships and in marriages. In the next Section 5 we shall undertake a complementary micro-founded analysis of the problem and show that, also at the level of individual strategies, our data on friendship nominations and marriages is consistent with our model.

We start our discussion in this section by defining the Coleman index, a measure to quantify the phenomenon of homophily. We highlight some stylized facts on the Coleman index that are observed in our data reflecting friendship nominations and marriages. Our analysis of homophily focuses on race, a significant characteristic along which distinct groups can be suitably defined. As we will discuss in greater detail in Section 5, friendship nominations plausibly reflect a one-sided matching scenario, while marriages fit well with a two-sided scenario. A detailed discussion of each of these data sets is provided in Section 5.[14]

### Measuring homophily

Recall from Subsection 2.1 the basic framework: there is a set $N \subset \mathbb{N}$ consisting of $n$ agents, who are partitioned into $q$ groups (or "types") indexed by $l$ and with respective cardinality $n_l$. We now denote by $w_l$ the relative population share of group $l$: $w_l \equiv \frac{n_l}{n}$. The measures of homophily we are about to discuss aim to quantify to what degree the type-distribution of matches is biased in favor

[14] Further details on data and empirical procedures are also available in Appendix 2.

of same-type matches. To this end, we denote by $m_{ll'}$ the number of matches between agents of type $l$ and agents of type $l'$, and by $m_l \equiv \sum_{l'=1}^{q} m_{ll'}$ the number of *total* matches of agents of type $l$.

A basic measure can be obtained by considering the ratio $\frac{m_{ll'}}{m_l}$, expressing the representation of type $l'$ matches in the total matches of group $l$. The particular case where $l' = l$ gives rise to what is called the **homophily index** of group $l$, $H_l \equiv \frac{m_{ll}}{m_l}$. This index is to be compared with the expected proportion of same-type matches that would result if matches resulted from a uniform random assortment process. Such a comparison is simply captured by what we shall call **excess homophily**, which is defined as the difference $H_l - w_l$ for each group $l$.[15]

When it comes to comparing the homophily of different groups the excess homophily index may provide a distorted picture. Groups with very large size $w_l$ will never experience large excess homophily as the maximal potential value of $H_l - w_l$, $1 - w_l$, is small. The index proposed by Coleman (1958) addresses the problem by normalizing the excess homophily of group $l$ by its maximal value $1 - w_l$. This gives rise to what we shall call the **Coleman (Homophily) Index**, which is defined as follows:

$$C_l = \frac{H_l - w_l}{1 - w_l}. \tag{8}$$

### Empirical patterns for the Coleman index

Much attention has been devoted to the relationship between a groups' tendency to inbreed and their relative population share. In Figure 1 we report plots of the Coleman homophily index against relative population shares for U.S. highschool friendships and U.S. marriages. Each observation corresponds to a particular racial group in a corresponding population. The left panel is similar to CJP: each dot refers to friendships for a specific ethnic group in a specific school. The right panel is new, and each dot refers to marriages for a specific ethnic group in a specific city-year combination (please refer to section 6 for a detailed description of the datasets).

In both plots, high values of the index are found consistently for groups that cover approximately half of the population.[16] The non-linear, hump shaped relationship between the Coleman index and relative population share was used by CJP to highlight the role of meeting biases in the process of friendship formation. In particular, positive levels of the index for all groups, and large positive values for middle sized groups, led to CJP's conclusion that some bias must be at work

[15] A positive difference between the index $H_l$ and the population share of group $l$ is usually referred to as "inbreeding homophily" of group $l$. We do not use this terminology here in order to avoid confusion with what we refer as the "inbreeding" choice of agents in our model.

[16] The main difference between the right and the left panels of Figure 1 is that in the right one for marriages the regressed values of the $C_q$ at zero and one are significantly different from zero. The intercept of the $C_q$ locus was used in Franz, Marsili and Pin (2008) to measure the bias in the meeting process.

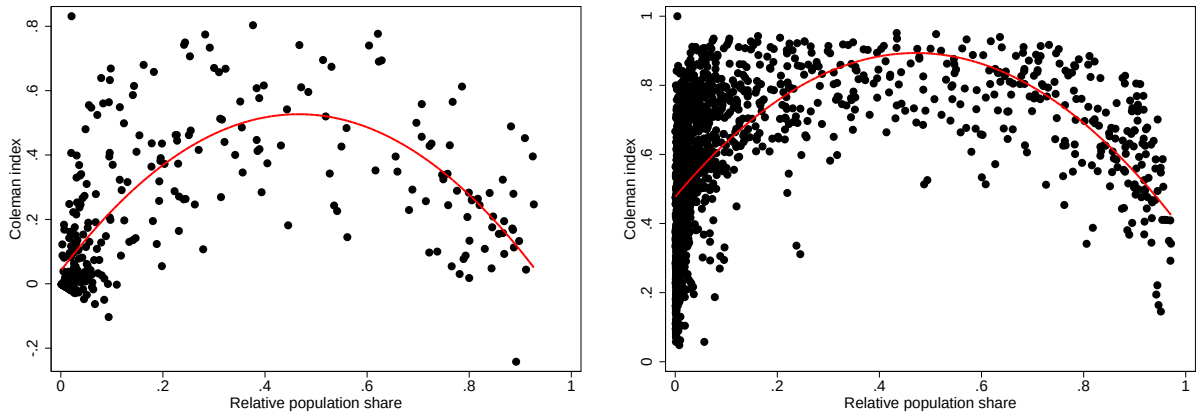


Figure 1: Coleman homophily index: Friendship nominations (left) and marriages (right).

in the meeting process. Indeed, were such biases absent—so that agents would meet uniformly at random—their model would imply that small groups should display negative values of the index, while for groups that comprise half of the population the index should approach zero.

### Matching the model to empirical patterns

The main qualitative features of Figure 1 can be summarised as:

1. The when population share is small the CI is small and positive. As population shares approach 0 friendships have, on average, a null CI and marriages have a strictly positive CI.
2. The CI is first increasing with population share.
3. The CI approaches 1 for groups with population shares close to half.
4. The CI decreases for groups with large population shares, becoming very small (and negative for friendships) as population share approaches 1.

We now match each of these features to the model, distinguishing between the one-sided and the two-sided scenarios.

First consider the two-sided scenario. The next proposition shows that, in this context, the expected CI's of outbreeding groups satisfy the following two properties: (a) they are strictly positive and bounded away from zero; (b) they are increasing in the groups' population shares. Note that both of these conclusions are in line with the evidence for marriages depicted in the right-hand panel of Figure 1, where the CI of very small groups is strictly positive and increasing. To state formally the result, let us denote by $\tilde{C}_l$ the ex-ante random variable that determines the CI of a group $l$ and by $\mathbb{E}[\tilde{C}_l]$ its expected value.

PROPOSITION **1** *Consider the two-sided scenario with $\hat{\tau}$ denoting the maximum equilibrium threshold and $n > 2q\hat{\tau}$. Let $l$ be any given outbreeding group. Then, $\mathbb{E}[\tilde{C}_l]$ is strictly positive and bounded away from zero, uniformly in the population size $n$. Furthermore, if $l$ and $l'$ are two outbreeding groups with $n_l < n_{l'}$, then $\mathbb{E}[\tilde{C}_l] < \mathbb{E}[\tilde{C}_{l'}]$.*

The one-sided scenario requires additional structure. For, in this context, the Expected Coleman Index (the term $\mathbb{E}[\tilde{C}_l]$, referred to as ECI henceforth in non-formal discussion) predicted at equilibrium for an outbreeding group approaches zero when the overall population grows large, as outbreeders randomly draw from the population at large. We next show, however, that by enriching the model with a small noise term (reflecting an element of pure randomness which affects agents' realized meetings) the increasing pattern that in Figure 1 applies to groups with small population shares (and hence outbreeding) also characterizes the one-sided scenario. Specifically, we posit:

**(F)** Independently of their breeding choice, all agents obtain a certain number $r_I > 1$ of draws from their own type as well as some number $r_O > 1$ from the population at large.

Since (F) is conceived as a "perturbation," the numbers $r_I$ and $r_O$ are to be thought as small. The best way to think of these noise terms is to imagine that not all realized meetings are under the full control of agents. In particular, even an inbreeder may end up meeting people from outside her group, possibly due to chance or to social or institutional constraints. And, similiarly, we also allow for the possibility that outbreeders direct a small part of their search within their own group only, be it for cultural, social, geographical or familial constraints. To repeat, however, these forces are thought as small.

The implications of such noise effects on the ECI of small outbreeding groups in the one-sided scenario are the object of the next proposition. First, it asserts that small outbreeding groups display a small ECI if frictions are small—in particular, if $r_I$ is small relative to $\eta$. Second, it indicates that the ECI is increasing in population shares for large enough populations.

PROPOSITION **2** *Consider the one-sided scenario under (F). There exists some $\hat{n}$ such that if $n \geq \hat{n}$, the following applies:*
*(i) Let $l$ be an outbreeding group $l$ of size $n_l$. Then $\mathbb{E}[\tilde{C}_l]$ is bounded above by $\frac{r_I}{\eta}$.*
*(ii) Let $l$, $l'$ be two outbreeding groups with $n_l < n_{l'}$. Then, $\mathbb{E}[C_l] < \mathbb{E}[C_{l'}]$.*

Our next result concerns groups of intermediate relative population share, which are inbreeding for large enough total population size. For these groups, as we next state formally, the ECI is arbitrarily high if meeting frictions are small, in both the one-sided and two-sided scenarios.

PROPOSITION **3** *Consider either the one- or the two-sided scenario under (F). Given any $\epsilon > 0$, there exist some positive $\delta_1$, $\delta_2$, $\delta_3$, and $\hat{n}$ such that if $n \geq \hat{n}$ and $\frac{r_O}{\eta} \leq \delta_3$ then any group $l$ with relative population share $\delta_1 > w_l > \delta_2$ has $\mathbb{E}[\tilde{C}_l] \geq 1 - \epsilon$.*

Finally, the next two results complete the present analysis by establishing how the homophily index changes with group size among relatively large (inbreeding) groups. First, Proposition 4 states that, among groups that inbreed and have a non-negligible relative population share, the expected Coleman index decreases as size grows. Second, Proposition 5 indicates that as a group approaches a situation of almost complete dominance (i.e. a fraction of the whole population that is close to one), its Coleman index falls to the point of becoming negative.

PROPOSITION **4** *Consider either the one- or the two-sided scenario under (F). Let $l$ and $l'$ be two groups whose relative population shares are bounded away from 0 and 1 (i.e. there exists some $\vartheta > 0$ such that $1 - \vartheta \geq w_{l'} > w_l \geq \vartheta$). Then, for* any *$\varpi > 0$, there exists some $\hat{n}$ such that if $n \geq \hat{n}$ and $w_{l'} - w_l \geq \varpi$, $\mathbb{E}[\tilde{C}_l] > \mathbb{E}[\tilde{C}_{l'}]$.*

PROPOSITION **5** *Consider either the one- or the two-sided scenario under (F). There exist some $\hat{n}$ and $\delta_1 > \delta_2 > 0$ such that if $n \geq \hat{n}$, then any group $l$ with relative population share $1 - \delta_2 \geq w_l \geq 1 - \delta_1$ has $\mathbb{E}[\tilde{C}_l] < 0$.*

We end this section by considering a remarkable implication of our threshold equilibrium for the patterns of the Coleman Index in small vs. large populations. This part is motivated by the observation, made by Currarini, Jackson and Pin (2010), that in the AddHealth dataset on high school friendships, school size (in terms of total number of students) significantly affects the homophilous (ethnic) bias in student friendships—see Figure 2. They report, in particular, that larger schools (those with more than 1000 students) display larger Coleman indices than smaller schools, which is a feature that their model can not directly accommodate. Intuitively, such an increase in homophily is in line with the main idea underlying our approach—i.e. that a minimal group size is needed for "inbreeding" activities to be effective. In fact, as we now argue, our theoretical setting provides a formal argument in support of these intuitions.[17]

Let $\tau$ be the equilibrium threshold, below which a group finds it profitable to outbreed. As it is shown in Theorems 1 and 2, this threshold size refers to the absolute number of agents in the group and is independent of the size of the network for large $n$. In particular, this threshold is *not* defined in terms of the relative population share of groups (that is, their fraction of the total population), which is measured on the horizontal axis of Figure 2. As the number $n$ of students in the school increases, there is a larger absolute size (that is, a larger number of group members)

[17] We are grateful to Matt Jackson for pointing out to us this property of our model.

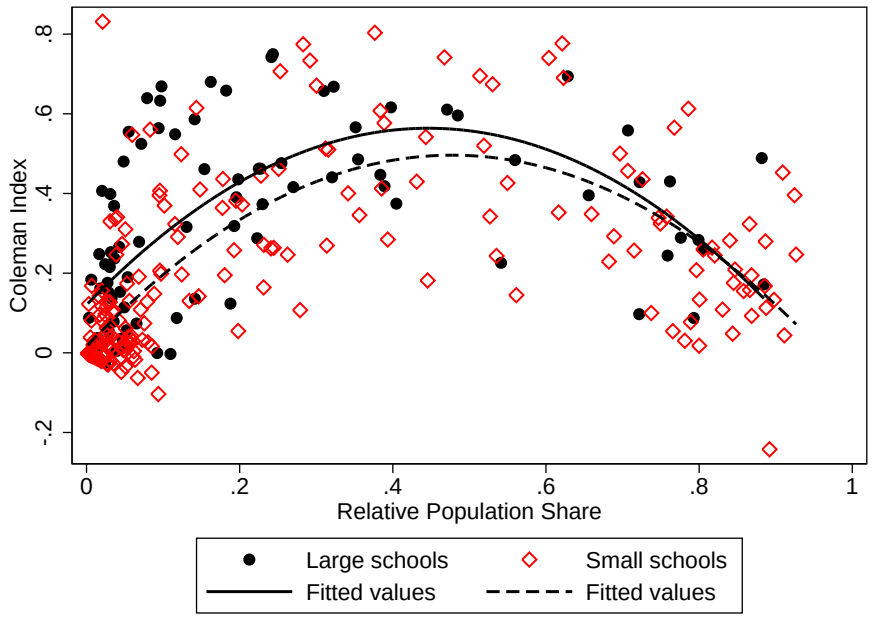


Figure 2: Coleman index in friendship nominations: Small schools (<1000) vs. large schools (>1000).

associated to any given group population share $w$. Thus denoting by $w(\tau, n)$ the relative population share that corresponds to the $\tau$ threshold for total population $n$, $w(\tau, n)$ is obviously decreasing in $n$. So, as population increases from $n$ to $n'$, those groups with population share $w$ such that $w(\tau, n') < w < w(\tau, n)$ start inbreeding and experience an increase in their Coleman index, while all other groups maintain their in/outbreeding strategy unaffected.

This logic can be used to explain the shift in the relation between Coleman index and relative population share that we observe in Figure 2. The shift is substantial for small and medium sized groups, and vanishes for very large groups. Indeed, in the sample of smaller schools, observations with small population share size are more likely to refer to groups with size below the threshold $\tau$. Therefore, as we shift attention to the sample of larger schools, we should expect to find a higher extent of inbreeding behavior. For observations corresponding to a medium population share, the increase in inbreeding is less significant since many of the observations among small schools correspond to groups that are already above the threshold. Finally, for observations with large relative population share, no significant change is observed because essentially all of those groups must be inbreeding, both in the sample of small and large schools.

# 5 Evidence from microlevel data

In Section 4 we demonstrate that the proposed theory replicates the aggregate behaviour of the Coleman homophily index. In this section we further examine some implications of the theory using microlevel data. The first implication concerns the strategic behaviour of individuals in both the one-sided and two-sided matching scenarios: a) Conditional on population share, individuals in large (absolute) groups will exhibit a greater tendency to inbreed than individuals in small (absolute) groups; b) Conditional on absolute group size, individuals in large (relative) groups will exhibit a greater tendency to inbreed than individuals in small (relative) groups. The microlevel data allow us to observe equilibrium matches, but not strategies *per se*. We show that, using the microlevel data, we can test a) but not b).

The second implication regards between-group matching in equilibrium, and differs between the two scenarios: Consider a small (absolute) group. Excess representation of other small groups will not significantly differ from zero when matching is one-sided, but excess representation of other small groups will be positive is matching is two-sided.

We utilise microlevel data reflecting friendship nominations and marriages. Friendship nomination provides a good example of one-sided matching. These data need not reflect consensual friendships; we find that fewer than 40% of all nominations are reciprocated. Therefore, friendship nominations provide useful information on one-sided matching, but should not be thought of as shedding light on consensual friendship formation. Clearly, marriages fit with a two-sided scenario, as both parties must agree to a observed match.

We emphasise that the purpose of this exercise is to test qualitative predictions of the model. Structural estimates of the primitives corresponding to the theory, which would be an interesting contribution, are beyond the scope of this paper.

## 5.1 Data

Here we provide an overview of the data used in our analysis. Summary statistics can be found in Appendix 2.

Friendship nomination data come from the Add health network structure files (Moody; 2005).[18] These files are constructed from the In-School questionnaire, administer across 90 118 adolescents, in grades 7–12 in the United States, for the 1994–1995 school year. These data are extensively used in sociological works on homophily (Moody (2001) for instance), and more recently by Currarini, Jackson and Pin (2009, 2010) in their economic model of friendship. In this questionnaire, students

[18]Data files are available from Add Health, Carolina Population Center (addhealth@unc.edu)

are asked to nominate up to 10 friends. The data record these friendship nominations and allow us to map networks of friendship nominations, by race, within a given school.

For the purpose of our analyse, we define four different racial categories in the Add Health data: White, Black, Hispanic and Asian. An observation is excluded if it does not include at least one observable nomination[19] or if the corresponding race is not identified by one of the four racial categories. The resulting sample includes 55 676 students across 78 different schools.

Marriage data come from four waves of the U.S. population census (Ruggles et al; 2015).[20] As with the friendship nominations, data should reflect a plausible link between match type (an in-group/out-group match is defined by race of husband relative to race of wife) and group characteristics (i.e. population share and absolute size). For this reason we focus the microlevel analysis on marital matches in cities identified in the public use population census.[21] We define five different racial categories: White, Black, American Indian (Native henceforth), Hispanic and Pacific Asian (Asian henceforth). The data include a linking rule that allows us to match spouses within a given census. Observations are excluded if: both spouses are not present in the data; wife does not belong to one of the five racial groups; the age of either spouse is less than 20 or greater than 49; either spouse is identified as immigrating after age 15. The first two conditions are required to identify the match-type. The third and fourth conditions increase the plausibility of the link between the match-type and group characteristics. This results in a sample of 501 235 marriages observed across 212 cities and 4 census-years.

We define the outcome of interest as follows. For an individual $i$ belonging to group $l$ in population $m$ the we define empirical variable $s^{\star}_{ilm}$, where $s^{\star}_{ilm} = 1$ if all observed matches are of race $l$ and $s^{\star}_{ilm} = 0$ if at least one match is of a different race than $l$. For the friendship nominations $i$, $l$ and $m$ correspond to student, race and school; a match corresponds to student $i$'s friendship nominations. For the marriage data $i$, $l$ and $m$ correspond to wife, race and city; a match corresponds to wife $i$'s husband. We drop subscripts for the remainder of the document.

## 5.2 Inbreeding and group size

The model implies that we should see a positive relationship between inbreeding and in-group size in both the one-sided matching and two-sided matching scenarios. Here we test this relationship using microlevel data.

[19] We only observe the racial characteristics of nominated students in the same school. For this reason, regressions weight observations by the proportion of total observations which we observe. For example, if a student has 7 nominations but only 2 are in that students school, they receive a weight of 2/7. This technique gives more weight to observations with more information.

[20] We utilize 5% samples for the years 1980, 1990 and 2000, and a 1% sample for 2010.

[21] For confidentiality purposes not all cities can be identified in the public files.

Let $s$ denote the true strategy played by an individual where $s = 1$ for *inbreed* and $s = 0$ for *outbreed*. Let $s^\star$ denote the empirical variable (defined above) where $s^\star = 1$ if all observed matches are *in-group* and $s^\star = 0$ if at least one observed match is *out-group*. We denote the probability of a random variable $x$ taking a specific value, $\bar{x}$, by $P[x = \bar{x}]$.

Causal observation suggests that $P[s^\star = 1]$ is increasing in $n_l$. In figures 3 and 4 we plot $P[s^\star = 1]$ against $n_l$ by race, for each school and city-year respectively[22]. For each group, in both the one-sided and two-sided scenarios, the proportion of observations matched only to in-group members is increasing. The increase in $P[s^\star = 1]$ is sharper at smaller sizes of the in-group.

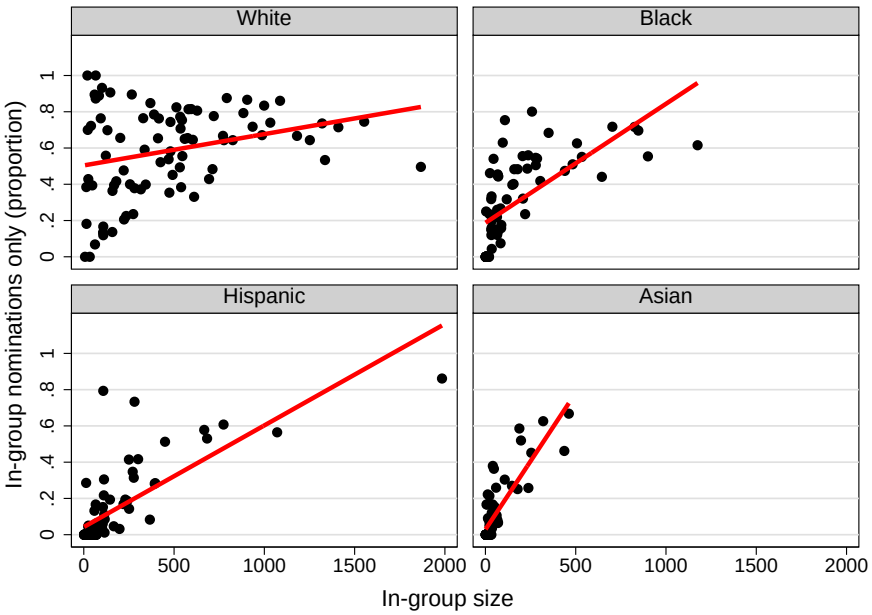


Figure 3: Proportion of students with all observed friendship nominations from in-group.

What can be learned about $s$ from the observation of $s^\star$? Recall: $\eta$ denotes the number of draws from the strategy-determined pool (either in-group or full population); $r_O > 1$ and $r_I > 1$ denote the number of draws, independent of strategy, made from the population, $N$, and the restricted in-group pool, $n_l$, respectively; $p$ denotes the probability of matching with any given draw. $\eta$, $p$, $r_O$ and $r_I$ can vary randomly across individuals but are independent of group size, population size and strategy. These parameters are not observed by the econometrician.

The relationship between $s^\star$ and $s$ can be described as follows:

$$s^\star = (1 - \mathbb{1}[\text{outgroup match}|r_O])s + (1 - \mathbb{1}[\text{outgroup match}|\eta + r_O])(1 - s), \tag{9}$$

[22]For presentation $n_l$ is restricted to less than 5 000 for cities. Figures do not qualitatively change is the full domain of values is considered.

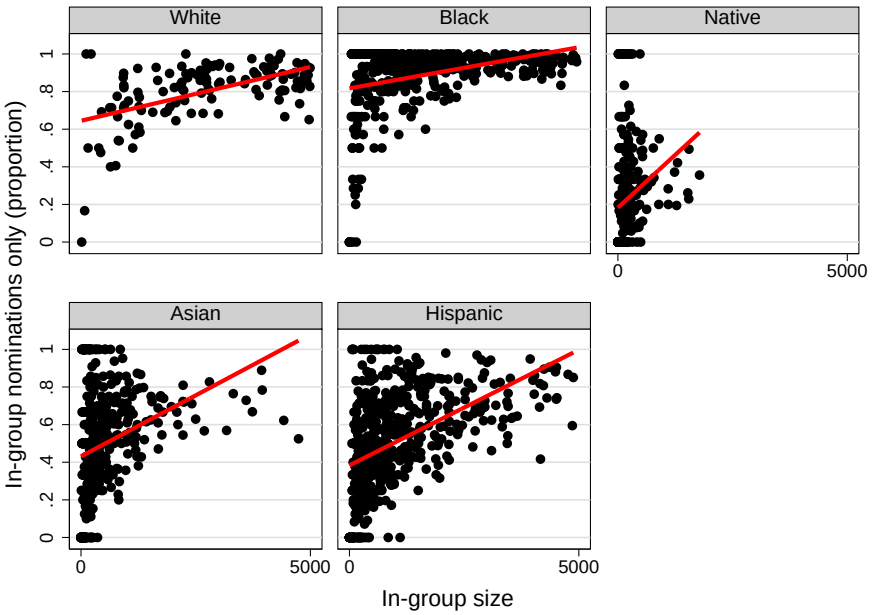


Figure 4: Proportion of marriages between in-group members.

where $\mathbb{1}[.|k]$ is an indicator function taking a value of 1 if the argument is true and 0 otherwise given $k$ draws from the full population. A couple of things are worth noting. First, when $s = 1$, we observe $s^\star = 1$ only if there is no *matching error*: $r_O$ draws do not yield a suitable out-group match. This is independent of $\eta$ and $r_I$ which are composed of draws only from the in group (and without error relative to the strategy). Second, when $s = 0$, we observe $s^\star = 0$ only if at least one of the $\eta + r_O$ draws is a suitable match from the out-group. This is independent of $r_I$ as these draws will never be from the out-group.

We show in Appendix 2 that the probability of $s^\star = 1$, given Equation 9, can be written as:

$$P[s^\star = 1] = P[s = 1][w_I + (1 - w_l)(1 - p)]^{r_O} + (1 - P[s = 1])[w_l + (1 - w_l)(1 - p)]^{\eta + r_O}$$

This equation highlights a fundamental problem with using $s^\star$ to draw conclusions about strategy. $P[s^\star = 1]$ is likely to provide a biased estimate of $P[s = 1]$. Further, the direction of this bias cannot be signed based on casual observation.

However, it is relatively straightforward to show that:

$$\frac{\partial P[s^\star = 1|w_l, n_l]}{\partial n_l} = \frac{\partial P[s = 1|w_l, n_l]}{\partial n_l}(\zeta^{r_O} - \zeta^{\eta + r_O}),$$

where $\zeta = w_l + (1 - w_l)(1 - p)$. Notice that $\zeta^{r_O} - \zeta^{\eta + r_O} > 0$; this implies that the sign of the observable right-hand-side behaviour of $s^\star$ is determined by the sign on the behaviour of $s$[23].

[23]However, the magnitude of the observed effect will under-estimate the magnitude of the strategy response. If

This has an important implication: observable matches ($s^\star$) can be used to test the qualitative relationship between $s$ and $n_l$.

A similar inference cannot be made from the behaviour of $\frac{\partial P[s^\star=1|w_l,n_l]}{\partial w_l}$. Intuitively, if a positive value is observed, we cannot determine whether this is due to an increase in $P[s^\star = 1]$, or whether it is due to an increase in the probability that out-breeders will only match with in-group members. Therefore, we turn to regression analysis and focus on the relationship between $P[s^\star = 1]$ and $n_l$.

We use regression analysis to test the positive relationship between $P[s = 1]$ and $n_l$, conditioning on $w_l$, implied by the model. We estimate a Probit model, for each race, regressing $s^\star_{ilm}$ on absolute size, $n_l$, and relative population share, $w_l$, of group $l$ (as well as the interaction between the two measures). In addition to group size, friendship nomination regressions include the dummy variables to control for student grade and marriage regressions include dummy variables for year of census, the age of each spouse and the education of each spouse. The estimated coefficients are reported in Table 1, for friendship nominations in the top panel and for marriages in the bottom panel.

The estimates reported in Table 1 are generally consistent with model. Two estimated coefficients corresponding to $n_l$—for *White* for friendship nominations and *Black* for marriages—are small in magnitude and statistically indistinguishable from 0.[24] For both friendship nominations and marriages we find: 1) the probability of observing $s^\star = 1$ is increasing with absolute in-group size; 2) the probability of of observing $s^\star = 1$ is increasing with relative population share; 3) the positive relationship between absolute in-group size and $s^\star$ is smaller when population share is large. Further 1), 2) and 3) are all consistent with the model. 2) is consistent with a low return to outbreeding (reflected by a small $w_l$ for a given $n_l$) leading to a higher probability of inbreeding. A possible interpretation of 3) can be presented as: When $w_l$ is close to 1 inbreeding is high; there are few individuals who will change strategy if $n_l$ where to increase. When $w_l$ is close to 0 outbreeding is high; there are many individuals who may change strategy if $n_l$ where to increase. While both 2) and 3) are consistent with the model, we are cautious about interpreting this as evidence, for the reasons discussed above. However, the estimates for $n_l$, in 1), provide strong support for the model.

We can use the friendship nomination data to further explore a consequence of the proposed model. In the model, the homophily observed in aggregate data results from the strategic choice of individuals to inbreed or outbreed. Consider an alternative theory: All individuals outbreed, $P[s = 1] = 0$, but they do so with an own-group matching bias. For example, this can be modelled as $p$ taking a greater value for in-group draws than for out-group draws. We can test whether this

values of $\eta$ and $r_O$ were observed we could derive the magnitude as well.

[24]Note that these are two of the largest groups in the samples. This may reflect the fact that these groups are generally of a size that an increase in $n_l$ has little impact on observed strategic choices (a large portion of each population is already playing an in-group strategy).

Table 1: Probit regression, one-sided and two-sided matching (outcome $s^\star$).

| | White | Black | Hispanic | Asian | Native |
|---|---|---|---|---|---|
| *One-sided*† *(Friendship nominations)* | | | | | |
| $n_l$ | 0.038 | 0.102 | 0.139 | 0.551 | |
| | (0.066) | (0.041)** | (0.037)*** | (0.062)*** | |
| $w_l$ | 2.231 | 2.012 | 2.872 | 3.883 | |
| | (0.369)*** | (0.200)*** | (0.357)*** | (1.440)*** | |
| $n_l \times w_l$ | -0.068 | -0.180 | -0.140 | -1.171 | |
| | (0.078) | (0.052)*** | (0.035)*** | (0.320)*** | |
| Obs. | 34 630 | 9 174 | 9 062 | 2 808 | |
| Pseudo $R^2$ | 0.061 | 0.069 | 0.276 | 0.159 | |
| *Two-sided*‡ *(Marriages)* | | | | | |
| $n_l$ | 0.052 | -0.016 | 0.832 | 0.139 | 7.001 |
| | (0.025)** | 0.022 | (0.073)*** | (0.024)*** | (2.039)*** |
| $w_l$ | 1.164 | 1.673 | 7.660 | 3.114 | 41.253 |
| | (0.113)*** | (0.217)*** | (0.768)*** | (0.252)*** | (6.786)*** |
| $n_l \times w_l$ | -0.073 | -0.092 | -5.526 | -0.427 | -353.237 |
| | (0.041)* | 0.084 | (0.586)*** | (0.089)*** | (65.285)*** |
| Obs. | 365 442 | 85 805 | 8 377 | 38 626 | 2 985 |
| Pseudo $R^2$ | 0.0456 | 0.0849 | 0.0977 | 0.1764 | 0.0667 |

Notes: Robust standard errors, clustered by school (one-sided) or city (two-sided), reported in parenthesis. *, **, and *** denote statistical significantly at 10%, 5% and 1%. Observations are weighted by the proportion of total nominations observed

† Additional co-variates include dummy variables for student's grade at time of survey. Data from Add Health Survey, see data appendix for details.

‡ Group size scaled by 10 000. Additional co-variates include dummy variables for: year of census, husband's and wife's age, husband's and wife's education.

alternative theory can explain observed patterns in the data, by looking at excess representation when $s^\star = 0$. If $p$ favours in-group draws then the representation of out-group matches relative to population share will be negative when $s^\star = 0$.

We calculate the *excess representation* , defined by:

$$\Delta_{ll'} \equiv \frac{m_{ll'}}{m_l} - w_{l'}. \tag{10}$$

Recall from Section 4 that $m_l \equiv \sum_{l'=1}^{q} m_{ll'}$ denotes the number of *total* matches of agents of type $l$. This measures the representation of a group $l'$ within the matches of group $l$ over and above what we would expect if all matches are made randomly. For example, we may record the number of Blacks, or Hispanics, or Asians among the total nominations reported by White individuals.

In Figure 5 we plot, for each racial group, out-group excess representation for the entire group (represented by the solid dots) and for the subgroups for whom $s^\star = 0$ (represented by hollow diamonds). The latter can be thought of as being largely composed of out-breeders. If there is an own-group bias in $p$, then we expect to see negative excess representation of outgroups when looking $s^\star = 0$. Figure 5 is summarized in Table 2 where we report, by race, the estimated mean (and standard error) of $\Delta_{ll'}$, for the full sample and conditional on $s^\star = 0$; we denote the estimated means by $E(\Delta_{ll'})$ and $E(\Delta_{ll'}|s^\star = 0)$ respectively.

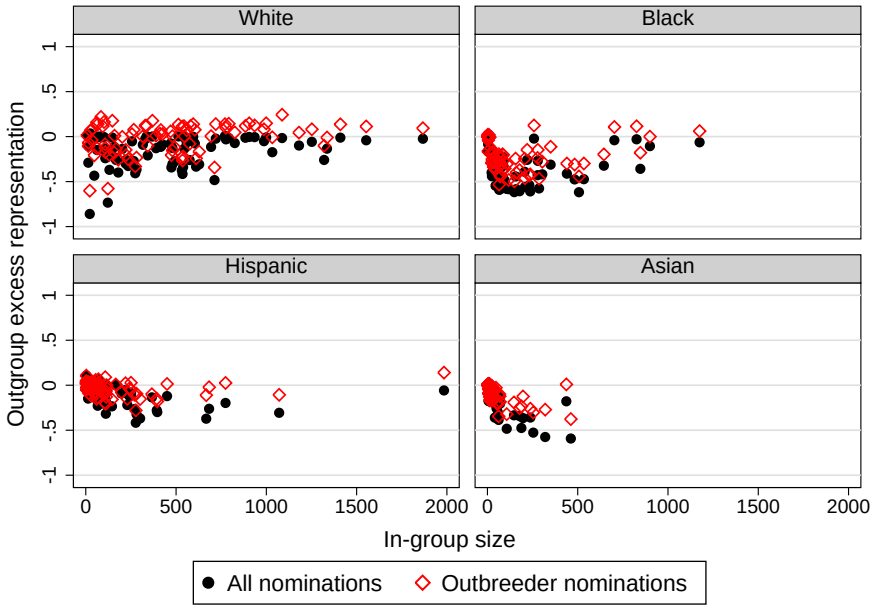


Figure 5: Excess representation for all students and $s^\star = 0$ only: Friendship nominations.

Both the fig. 5 and the mean values reported in Table 2 suggest that the alternative theory cannot fully explain the observed patterns in friendship nomination. This is particularly stark for White and Hispanic groups for which $\Delta_{ll'}$ does not significantly differ from 0 in the $s^\star = 0$ subgroup.

### 5.3 Cross-group ties

Which scenario we are in—one-sided matching versus two-sided matching— will have implications for the predictions of our model with respect to the cross-group ties observed in equilibrium. The

Table 2: Excess representation for all students and $s^\star = 0$ only: Friendship nominations.

| | White | Black | Hispanic | Asian |
|---|---|---|---|---|
| $E(\Delta_{ll'})$ | -0.119 | -0.316 | -0.173 | -0.342 |
| | (0.017)*** | (0.044)*** | (0.028)*** | (0.047)*** |
| $E(\Delta_{ll'}\|s^\star = 0)$ | 0.006 | -0.192 | -0.045 | -0.195 |
| | (0.017) | (0.041)*** | (0.030) | (0.0377)*** |

Notes: Robust standard errors in parenthesis. *, **, and *** denote excess representation significantly differs from 1 at 10%, 5% and 1%.

two-sided scenario predicts that outbreeders meet agents in the restricted pool of outbreeders. Thus, if meeting is uniform, outbreeding groups should display an excess representation of other outbreeders. This simply follows from the fact that outbreeding groups are found with probabilities that reflect the relative shares *in the pool of outbreeders*, and these shares exceed those in the overall population. Since outbreeding groups are relatively small, it follows that cross-group matches are primarily formed among agents of small groups.

To examine this prediction we look at the relative representation of small groups in friendship nominations and the marriages of other small groups.[25] Figure 6 plots excess representation of small groups (racial groups with fewer than 80 members in a school (nominations) and 500 members in a city(marriages)) against own-group size (restricted to groups smaller than 5 000). Formally, for each group $l$, we calculate $\sum_{\{l':n_{l'}\leq x\}} \Delta_{ll'}$, where $x$ stands for the corresponding threshold ($x = 80$

[25] As very small groups are relatively few in number, we do not analysis separately by race.

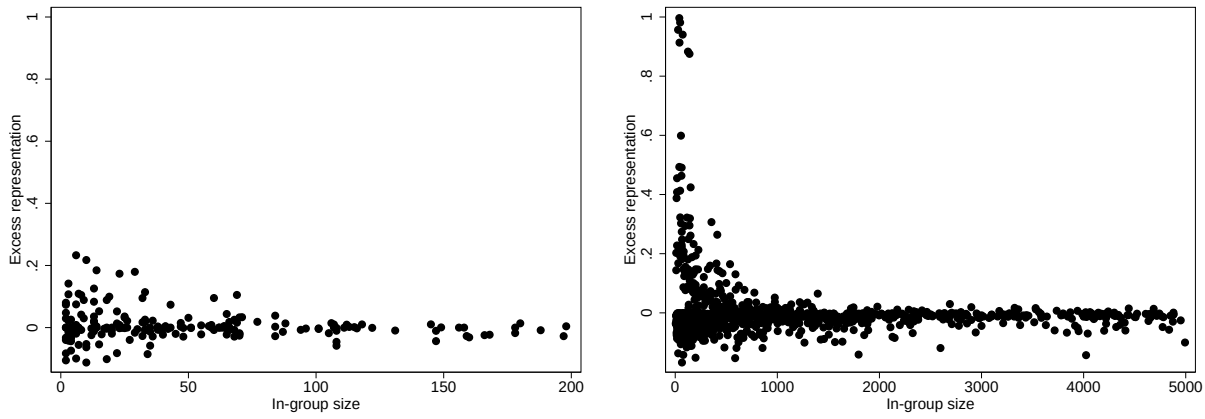


Figure 6: Excess representation of small groups: Friendship nominations (left) and marriages (right).

Table 3: Excess representation of small groups.

| *One-sided (Friendship nominations)* | In-group size | | |
|---|---|---|---|
| | 1–20 | 20–50 | 50–200 |
| $E(\sum \Delta_{ll'} \mid n_{l'} \leq 80)$ | 0.013 | 0.008 | -0.000 |
| | (0.009) | (0.008) | (0.004) |
| Number of schools | 68 | 51 | 56 |
| *Two-sided (Marriages)* | | | |
| | 1–50 | 50–200 | 200–1 000 |
| $E(\sum \Delta_{ll'} \mid n_{l'} \leq 500)$ | 0.048 | 0.007 | -0.006 |
| | (0.018)*** | (0.006) | (0.002)** |
| Number of cities | 112 | 148 | 155 |

Notes: Values reflect means. Robust standard errors, clustered by school (nominations) and city (marriages), reported in parenthesis. *, **, and *** denote excess representation statistically differs from 0 at 10%, 5% and 1%.

for nominations and $x = 500$ for marriages). These cases are shown for illustrative purpose, and qualitatively similar pictures obtain when we fix different small thresholds.

Figure 6 suggests that a positive excess representation of "small" groups is a feature of the marriages (right-hand panel) of small groups only, while for larger groups tend to marry with these small groups at rates below these groups' population shares. In particular, there seems to be some very small critical size of groups after which the over-representation of small groups disappears.

This insight is supported in Table 3, where we report the mean excess representation of small groups by different racial groups, denoted by $E(\sum \Delta_{ll'} \mid n_{l'} \leq x)$, stratifying, for each $l$, by in-group size from 1–20, 20–50 and 50–200 for friendship nominations and 1–50, 50–200 and 200–1 000 for marriages[26]. Consistent with our theory, mean excess representation is significant and positive in two-sided matching, but not distinguishable from zero in the one-sided matching. In the two-sided scenario, mean excess representation decreases for medium sized in-groups and is significant and negative for large in-groups. These results are consistent with the predictions of the model for two-sided and one-sided matching.

[26]Bin sizes are chosen to keep the number of schools/cities relatively constant.

## 6 Summary and concluding remarks

The paper has proposed a very stylized model of homophily, which may be applied to a diverse range of alternative phenomena such as friendships and marriages. Our main purpose has been to provide a behavioral foundation to the meeting biases that have been shown in previous works to play a key role in the emergence of homophily in social networks. Our approach hinges upon two key assumptions: (i) the establishment of ties with individuals that differ in some relevant characteristics (e.g. race or language) implies a costly investment; (ii) the search for suitable ties is more effective in larger pools. Under these assumptions, the induced game was shown to have a threshold equilibrium where groups outbreed if, and only if, their size falls below a certain level. This simple structure of the equilibrium has implications that match the empirical evidence found in both friendship and marriage data. Specifically, it is consistent with the nonmonotonicity displayed by the Coleman homophily index as well as with regularities observed on the pattern of in-group and cross-group ties.

While homophily is a complex and multifaceted phenomenon, we believe that our model highlights a very basic force underlying homophily that future analysis of the phenomenon may take into account. Our extremely stylized model does not contain explicit elements of preferences, and homophily is built in through the (fixed) cost of outbreeding. A more realistic model allowing for preferences would contain additional interesting features, that we plan to integrate in the present framework in future research. In fact, in the presence of preferences in favour of in-group contacts, groups accounting for a small share of the population may face strong incentives to inbreed in order to avoid mixes of realized contacts dominated by the out-group. Other issues that future research should address include the consideration of flexible individual characteristics. In many social contexts, these characteristics (language, religion, etc.) are not forever fixed in individuals and their descendants but can be changed through inter- action which may possibly mitigate differences, but also exacerbate them in some other cases. In this sense, cross-ties among different types could breed convergence of characteristics (and thus integration), or possibly the opposite. In general, one might anticipate that interesting nonlinear dynamics may arise under some circumstances. To understand better such interplay between in- teraction/segmentation on the one hand and homogenization/polarization on the other, seems a crucial issue for future theoretical and empirical research.

## References

[1] Allport, W. G., 1954. *The Nature of Prejudice*. Cambridge, MA: Addison-Wesley.

[2] Aubrun, G. and I. Nechita, 2009. "Stochastic Ordering for Iterated Convolutions and Catalytic Majorizations," *Annales de l'Institut Henri Poincaré— Probabilité et Statistiques* 45(3), 611–625.

[3] Baccara, M. and L. Yariv, 2013. "Homophily in Peer Groups," *American Economic Journal: Microeconomics* 5(3), 69–96.

[4] Bala, V. and S. Goyal, 2000. "A Noncooperative Model of Network Formation," *Econometrica* 68(5), 1181–1229.

[5] Blau, P. M., 1977. *Inequality and Heterogeneity: A Primitive Theory of Social Structure*. New York: Free Press.

[6] Bramoullé, Y. and B. Rogers, 2009. "Diversity and Popularity in Social Networks," mimeo.

[7] Coleman, J., 1958. "Relational Analysis: The Study of Social Organizations With Survey Methods," *Human Organization* 17, 28–36.

[8] Currarini, S., M.O. Jackson and P. Pin, 2009. "An Economic Model of Friendship: Homophily, Minorities and Segregation," *Econometrica* 77(4), 1003–1045.

[9] Currarini, S., M.O. Jackson, and P. Pin, 2010. "Identifying the Roles of Choice and Chance in Network Formation: Racial Biases in High School Friendships," *Proceedings of the National Academy of Science* 107, 4857–4861.

[10] Dixit, A., 2003. "Trade Expansion and Contract Enforcement," *Journal of Political Economy* 111(6), 1293–1317.

[11] Fischer, CS , 1982, *To Dwell among Friends*, Chicago, Univ. Chicago Press.

[12] Franz, S., M. Marsili and P. Pin, 2008. "Observed choices and underlying opportunities," *Science and Culture* 76, 471–476.

[13] Galeotti, A., F. Ghiglino and F. Squintani, 2013. "Strategic Information Transmission Networks" *Journal of Economic Theory* 148(5), 1751–1769.

[14] Giles, M. W., 1978. "White Enrolment Stability and School Desegregation: A Two-Level Analysis," *American Sociological Review* 43, 2448–2464.

[15] Golub, B. and M. O. Jackson, 2012. "How Homophily Affects the Speed of Learning and Best-Response Dynamics," *Quarterly Journal of Economics* 127(3), 1287–1338.

[16] Jackson, M. and B. Rogers, 2005. "The Economics of Small Worlds," *The Journal of the European Economic Association (papers and proceedings)* 3(2-3), 617–627.

[17] Kandel, D., 1978. "Homophily, Selection, and Socialization in Adolescent Friendships," *American Journal of Sociology* 84(2), 427–436.

[18] Lazarsfeld, P.F. and R.K. Merton, 1954. "Friendship as a Social Process: A Substantive and Methodological Analysis," in M. Berger (ed.), *Freedom and Control in Modern Society*, New York: Van Nostrand.

[19] Marsden, P.V., 1987. "Core Discussion Networks of Americans," *American Sociological Review* 52, 122–313.

[20] Marsden, P.V., 1988. "Homogeneity in Confiding Relations," *Social Networks* 10, 57–76.

[21] McPherson, M., L. Smith-Lovin and J. M. Cook, 2001. "Birds of a Feather: Homophily in Social Networks," *Annual Review Sociology* 27, 415–444.

[22] Moody, J., 2001. "Race, School Integration, and Friendship Segregation in America," *The American Journal of Sociology* 107(3), 679–716.

[23] Moody, J., 2005. "Add Health Network Structure Files," technical document: Carolina Population Center University of North Carolina at Chapel Hill.

[24] van der Poel, M. 1993. *Personal Networks*. Netherlands: Swets & B. V. Zeitlinger.

[25] Ruggles, S., K. Genadek, R. Goeken, J. Grover and M. Sobek, 2015. Integrated Public Use Microdata Series: Version 6.0 [Machine-readable database]. Minneapolis: University of Minnesota.

[26] Staje, W., 1990. "The Collector's Problem with Group Drawings," *Advances in Applied Probability* 22(4), 866–882.

[27] Suen, W, 2010. "Mutual Admiration Clubs," *Economic Inquiry* 48(1), 123–132.

[28] Vega-Redondo, F., 2007. *Complex Social Networks*, Econometric Society Monograph Series, Cambridge: Cambridge University Press.

[29] Zeggelink, E. P. H. 1993. *Strangers into Friends: The Evolution of Friendship Networks Using an Individual Oriented Modeling Approach*. Amsterdam: ICS

# Appendix 1

Here, we provide the proof for the formal results stated in the main text.

**Proof of Theorem 1**

First we note that, in the one-sided model, the payoff of any player $i$ belonging to an outbreeding group $l$ in a group-symmetric profile $\gamma$ is independent of the choice of groups different from $l$. Specifically, the expected payoff $\pi_l(\gamma)$ for an individual $i$ of an outbreeding group $l$ is given by the expression:

$$\pi_O(n_l) \equiv V(\eta) - c - \delta(n), \tag{11}$$

where $\delta(n) \to 0$ as $n \to \infty$. Similarly, we can write the payoff of any individual of group $l$ when inbreeding as:

$$\pi_I(n_l) \equiv \sum_{\nu_i=0}^{n} P_{\eta,n_l}(\nu_i)\ V(\nu_i). \tag{12}$$

Take the extreme case where $n_l = 1$. Obviously, $\pi_I(n_l) = 0$ while, by virtue of the assumption that

$$V(1) > c.$$

we have $\pi_O(1) > 0$. This implies that outbreeding is always optimal for sufficiently small groups.

Next, we want to show that such inbreeding incentives decrease monotonically with group size. To this end, we can invoke Lemma 1, already stated in Section 3, which claims that, as the pool size becomes larger, the induced distributions over the number of distinct meetings improve in the FOSD sense. Before proceeding with the proof of the Theorem, we provide a detailed proof of that auxiliary result.

*Proof of Lemma 1:*

Let us denote by $\overrightarrow{P}_\theta(\nu;\eta,)$ the probability of $\nu$ distinct elements from $\eta$ draws with replacement out of a set of size $\theta$. It is enough to show that, for all $\eta$ and $\theta$, the probability distribution $\left\{\overrightarrow{P}_{\theta+1}(\nu;\eta)\right\}_{\nu=0,1,2,...}$ dominates the distribution $\left\{\overrightarrow{P}_\theta(\nu;\eta)\right\}_{\nu=0,1,2,...}$ in the FOSD sense.

Following Staje (1990), we can write:

$$\overrightarrow{P}_\theta(\nu;\eta) = \binom{\theta}{\nu} \sum_{j=0}^{\nu} (-1)^j \binom{\nu}{j} \left(\frac{\nu - j}{\theta}\right)^\eta$$

Let us now consider the ratio of $\overrightarrow{P}_{\theta}(\nu;\eta)$ to $\overrightarrow{P}_{\theta+1}(\nu;\eta)$ :

$$\frac{\overrightarrow{P}_{\theta}(\nu;\eta)}{\overrightarrow{P}_{\theta+1}(\nu;\eta)} = \frac{\binom{\theta}{\nu}\sum_{j=0}^{\nu}(-1)^j\binom{\nu}{j}\left(\frac{\nu-j}{\theta}\right)^{\eta}}{\binom{\theta+1}{\nu}\sum_{j=0}^{\nu}(-1)^j\binom{\nu}{j}\left(\frac{\nu-j}{\theta+1}\right)^{\eta}} \tag{13}$$

which can be written as:

$$\frac{\frac{1}{\theta^{\eta}}\frac{\theta!}{(\theta-\nu)!\,\nu!}\sum_{j=0}^{\nu}(-1)^j\binom{\nu}{j}(\nu-j)^{\eta}}{\frac{1}{(\theta+1)^{\eta}}\frac{(\theta+1)!}{(\theta+1-\nu)!\,\nu!}\sum_{j=0}^{\nu}(-1)^j\binom{\nu}{j}(\nu-j)^{\eta}}$$

or, equivalently:

$$\frac{\frac{1}{\theta^{\eta}}\frac{\theta!}{(\theta-\nu)!\,\nu!}}{\frac{1}{(\theta+1)^{\eta}}\frac{(\theta+1)!}{(\theta+1-\nu)!\,\nu!}} = \frac{(\theta+1)^{\eta-1}(\theta+1-\nu)}{\theta^{\eta}}.$$

Note that for $\nu = 1$ this yields:

$$\frac{(\theta+1)^{\eta-1}}{\theta^{\eta-1}} > 1.$$

Note also that for all admissible values of $\theta$ and $\nu$, the ratio $\frac{\overrightarrow{P}_{\theta}(\nu;\eta)}{\overrightarrow{P}_{\theta+1}(\nu;\eta)}$ is decreasing in $\nu$. Since these are probability distributions, we conclude that these exists $\bar{\nu}$ such that $\frac{\overrightarrow{P}_{\theta}(\nu;\eta)}{\overrightarrow{P}_{\theta+1}(\nu;\eta)} < 1$ for all $\nu > \bar{\nu}$. This implies that $\overrightarrow{P}_{\theta+1}(\nu;\eta)$ First Order Stochastic Dominates $\overrightarrow{P}_{\theta}(\nu;\eta)$, and thus completes the proof of the Lemma.

Returning now to the proof of the Theorem, recall that $U(y_i+1) \geq U(y_i)$ for all $y_i \geq 1$ and $U(1) > U(0)$. Therefore, combining Lemma 1 with the monotonicity of $U$ we may conclude that, for any group size $n_l$,

$$\pi_I(n_l+1) - \pi_I(n_l) > 0. \tag{14}$$

Let now $\tau^*$ be the lowest integer such that

$$\pi_I(\tau) \geq V(\eta) - c. \tag{15}$$

Then, both if (15) holds strictly or with equality, it is clear that by making $n$ large enough, we have

$$\pi_I(\tau-1) < \pi_O(\tau) < \pi_I(\tau),$$

which proves that $\tau^*$ is the desired threshold, and completes the proof of the Theorem. ■

**Proof of Theorem 2**

In the present two-sided scenario, we find again that the threshold features of the equilibria hinge upon the monotonically decreasing incentives to outbreeding resulting from increasing pool size. Such monotonicity is the essential implication of Lemma 2, already stated in Section 3, which claims that the *expected number* of distinct meeting grows with pool size. Before tackling the proof of the Theorem itself, we provide a detailed proof of that Lemma.

*Proof of Lemma 2*

Given a set $\Theta$ and some $L \subset \Theta$, let us first derive (cf. Stadje (1990)) the expected number of distinct meetings that agent $i$ obtains from the set $\Theta \backslash L$ by means of $\eta$ independent draws with replacement out of the set $\Theta$. Denoting by $\theta$ and $l$, the cardinalities of the sets $\Theta$ and $L$ respectively, that expected number is equal to

$$(\theta - l) \cdot q_\theta(\eta) \tag{16}$$

where

$$q_\theta(\eta) = \left(1 - \left(\frac{\theta - 1}{\theta}\right)^\eta\right) \tag{17}$$

is the probability that an agent in the set $\Theta$ is found by means of $\eta$ draws with replacement from that set. (In our two-sided scenario, the set $L$ is to be interpreted as the set of agents that find $i$ through search, and that should not be counted twice in the union of passive and active draws if found also by agent $i$.)

Consider now the random variable $\tilde{\nu}_{\leftrightarrow}(\theta)$ considered in the statement of the Lemma, for some given pool size $\theta \in \mathbb{N}$. Recall that this variable gives the number of distinct meetings an agent obtains from a pool of size $\theta$ when meeting is two-sided and both this agent and all the others obtain $\eta$ draws. Its expected value can be computed by adding the expected number of agents who are met "passively" by this agent, i.e.

$$\sum_{l=0}^{\theta} \binom{\theta}{l} q_\theta(\eta)^l (1 - q_\theta(\eta))^{\theta - l} \cdot l$$

and those that are found through "active" search , i.e.

$$\sum_{l=0}^{\theta} \binom{\theta}{l} q_\theta(\eta)^l (1 - q_\theta(\eta))^{\theta - l} \cdot (\theta - l) \cdot q_\theta(\eta).$$

Thus, combining both expressions, we can write:

$$\mathbb{E}\left[\tilde{\nu}_{\leftrightarrow}(\theta)\right] = \sum_{l=0}^{\theta} \binom{\theta}{l} q_\theta(\eta)^l (1 - q_\theta(\eta))^{\theta - l} \cdot l + \sum_{l=0}^{\theta} \binom{\theta}{l} q_\theta(\eta)^l (1 - q_\theta(\eta))^{\theta - l} \cdot (\theta - l) \cdot q_\theta(\eta). \tag{18}$$

Now note that by factoring the term $\theta \cdot q_\theta(\eta)$ in the second summatory in (18) we can write this sum as follows:

$$\theta \cdot q_\theta(\eta) \sum_{l=0}^{\theta} \binom{\theta}{l} q_\theta(\eta)^l (1 - q_\theta(\eta))^{\theta - l} - q_\theta(\eta) \sum_{l=0}^{\theta} \binom{\theta}{l} q_\theta(\eta)^l (1 - p(\eta, \theta))^{\theta - l} l. \tag{19}$$

Note that the second term of (19) is just $\theta \cdot q_\theta(\eta)^2$, while the first term is simply $\theta \cdot q_\theta(\eta)$. Integrating all the former considerations into (18) we can write:

$$\mathbb{E}\left[\tilde{\nu}_{\leftrightarrow}(\theta)\right] = q_\theta(\eta) \cdot \theta + q_\theta(\eta) \cdot \theta - q_\theta(\eta)^2 \cdot \theta$$

Let us define the function $f(\theta, \eta)$ by the right-hand side of the above expression. The derivative of $f$ with respect to $\theta$ is given by:

$$\frac{\partial f(\theta, \eta)}{\partial \theta} = \frac{1}{\theta - 1} \left[ (\theta - 1) \left( 1 - \left( \frac{\theta - 1}{\theta} \right)^{2\eta} \right) - 2\eta \left( \frac{\theta - 1}{\theta} \right)^{2\eta} \right]$$

and the sign of $\dfrac{\partial f(\theta, \eta)}{\partial \theta}$ is the sign of the following expression:

$$(\theta - 1) \left( 1 - \left( \frac{\theta - 1}{\theta} \right)^{2\eta} \right) - 2\eta \left( 1 - \left( \frac{\theta - 1}{\theta} \right)^{2\eta} \right).$$

Taking logs we have that $\frac{\partial f(\theta, \eta)}{\partial \theta} > 0$ iff:

$$\ln(\theta - 1) > 2\eta \ln(\theta - 1) - 2\eta \ln(\theta) + \ln(2\eta + \theta - 1)$$

which rewrites as follows:

$$2\eta \left( \ln(\theta) - \ln(\theta - 1) \right) > \ln(2\eta - 1 + \theta) - \ln(\theta - 1).$$

The above condition is a direct consequence of the strict concavity of the logarithm function, which establishes the Lemma.

Under the assumption that the utility function is linear, Lemma 2 readily implies that, in every group-symmetric Nash equilibrium of the breeding game, if an agent of group $l$ inbreeds then every other individual of a group $l'$ such that $n_{l'} > n_l$ must inbreed as well. The equilibrium, therefore, must be of the threshold type. Finally, we argue that one such equilibrium is defined by the same threshold $\tau^*$ established in Theorem 1 that defines the (unique) equilibrium in the one-sided scenario. To see this, simply note that, if $\sum_{l : n_l < \tau *} n_l > \alpha n$ we can still write

$$\pi_O(n_l) \equiv V(\eta) - c - \delta(n), \tag{20}$$

where $\delta(n) \to 0$ as $n \to \infty$. Therefore, for large enough $n$ the same argument as in the proof of Theorem 1 establishes $\tau^*$ as the only equilibrium threshold in this case, which completes the proof of the result. ■

**Proof of Proposition 1** First, consider any given outbreeding group $l$. Its expected excess homophily is $\mathbb{E}\left[\tilde{H}_l\right] - w_l = \frac{w_l - 1/n}{w_O - 1/n} - w_l$, where $w_O$ denotes the equilibrium fraction of outbreeders in the population. If $n > \hat{\tau}\, q$, since $n_O \leq q\hat{\tau}$, then

$$\frac{w_l - 1/n}{w_O - 1/n} = \frac{n_l - 1}{n_O - 1} - \frac{n_l}{n} \geq \delta$$

for some $\delta > 0$, uniformly in $n$. Therefore, we have that

$$\mathbb{E}\left[\tilde{C}_l\right] = \frac{\mathbb{E}\left[\tilde{H}_l\right] - w_l}{1 - w_l} \geq \delta > 0 \tag{21}$$

as claimed in the first part of the proposition. As for its second part, it immediately follows from the fact that, in the expression for $\mathbb{E}\left[\tilde{C}_l\right]$, the numerator is increasing with $w_l$ while the denominator decreases with it, for given $w_O$.

**Proof of Proposition 2**

(i) Consider an outbreeding group $l$ of given size $n_l$. First note that, as $n \to \infty$, we have $w_l = \frac{n_l}{n} \searrow 0$. Thus, for $n$ large enough, the random variable $\tilde{C}_l$ can be approximated as follows:

$$\tilde{C}_l \simeq \frac{\tilde{\nu}(r_I, n_l)}{\tilde{\nu}(r_I, n_l) + \tilde{\nu}(\eta + r_O, \infty)}$$

and, therefore, for some arbitrarily small $\epsilon$, one can write:

$$\begin{aligned}
\tilde{C}_l &\leq \frac{r_I}{\eta + r_O + r_I} + \epsilon = \frac{r_I}{\eta}\left(1 - \frac{\eta}{\eta + r_O + r_I}\right) + \frac{\epsilon}{\eta} \\
&= \frac{r_I}{\eta} - \frac{r_I}{\eta}\left(1 - \frac{\eta}{\eta + r_O + r_I}\right) + \frac{\epsilon}{\eta} \\
&\leq \frac{r_I}{\eta}
\end{aligned}$$

if $\epsilon$ is chosen small enough.

(ii) For simplicity, consider two outbreeding groups $l$ and $l'$ whose cardinalities differ in just one individual, i.e. $n_l + 1 = n_{l'}$, and let $\Delta \equiv \left\{\mathbb{E}\left[\tilde{H}_{l'}\right] - w_{l'}\right\} - \left\{\mathbb{E}\left[\tilde{H}_l\right] - w_l\right\}$ denote the expected change in excess homophily. Furthermore, let $m(x, y) \equiv \mathbb{E}[\tilde{\nu}(x, y)]$ stand for the expected number of distinct meetings obtained when the number of draws is $x$ and the pool size is $y$. Then, we can

write:

$$\begin{aligned}\Delta &= \left[\frac{m(r_I, n_l')}{m(r_I, n_l') + m(r_O + \eta, \infty)} - \frac{n_{l'}}{n}\right] - \left[\frac{m(r_I, n_l)}{m(x, n_l) + m(r_O + \eta, \infty)} - \frac{n_l}{n}\right] \\ &= \frac{m(r_I, n_l')}{m(r_I, n_l') + m(r_O + \eta, \infty)} - \frac{m(r_I, n_l)}{m(r_I, n_l) + m(r_O + \eta, \infty)} - \frac{1}{n}.\end{aligned} \tag{22}$$

Since, by Lemma 2, $m(r_I, n_l') - m(r_I, n_l) > 0$, the difference

$$\frac{m(r_I, n_l')}{m(r_I, n_l') + m(r_O + \eta, n_l')} - \frac{m(r_I, n_l)}{m(r_I, n_l) + m(r_O + \eta, n_l)}$$

is strictly positive and uniformly bounded away from zero. It follows, therefore, that, for $n$ large enough, $\Delta$ is strictly positive. Recalling now the expression for the Coleman index, and since, obviously, $1 - w_{l'} < 1 - w_l$, the fact that $\Delta$ is strictly positive implies that $\mathbb{E}[C_l] < \mathbb{E}[C_{l'}]$ ■

**Proof of Proposition 3** A preliminary observation is that, if $n$ is large enough, then since $w_l = \frac{n_l}{n}$ is bounded away from zero by $\delta_2$ it must be that $n_l \geq \tau^*$ (where $\tau^*$ is as in Theorem 1) and therefore group $l$ must inbreed in any equilibrium, either in the one- or two-sided scenarios. The same considerations indicate that, for large enough $n$, the group size $n_l$ can be made arbitrarily large, in which case we can approximate its expected Coleman index as follows:

$$\mathbb{E}[\tilde{C}_l] \simeq \frac{\frac{m(\eta + r_I, \infty)}{m(\eta + r_I, \infty) + m(r_O, \infty)} - \frac{n_l}{n}}{1 - \frac{n_l}{n}} = \frac{\frac{\eta + r_I}{\eta + r_I + r_O} - w_l}{1 - w_l}.$$

An appropriate choices of $\delta_3$ ensures that the term $\frac{\eta + r_I}{\eta + r_I + r_O}$ is arbitrarily close to 1. Thus, by choosing $\delta_1$ small enough, the expected homophily $\mathbb{E}[\tilde{C}_l]$ can be made arbitrarily close to 1, as desired. ■

**Proof of Proposition 4** Consider two groups, $l$ and $l'$, whose relative population shares are bounded away from 0 and 1, as formulated in the statement of the result. As $n$ becomes large, both groups must exceed the threshold $\tau^*$ specified in Theorem 1, so both find it optimal to inbreed. Then, by invoking the usual approximations for large $n$ to approximate the expected Coleman index, the desired conclusion reads:

$$\frac{\frac{m(\eta + r_I, \infty)}{m(\eta + r_I, \infty) + m(r_O, \infty)} - w_l}{1 - w_l} > \frac{\frac{m(\eta + r_I, \infty)}{m(\eta + r_I, \infty) + m(r_O, \infty)} - w_{l'}}{1 - w_{l'}} \tag{23}$$

where we use the fact that the size of both groups grows unboundedly with $n$. In view of the fact that

$$\frac{m(\eta + r_I, \infty)}{m(\eta + r_I, \infty) + m(r_O, \infty)} = \frac{\eta + r_I}{\eta + r_I + r_O} < 1,$$

it is immediate to see that (23) holds if, and only if, the difference $w_{l'} - w_l$ is bounded above zero, as claimed. ■

**Proof of Proposition 5** Given $\eta$, $r_O$, and $r_I$, choose $\delta_1 < \frac{1}{2}\frac{m(r_O,\infty)}{m(\eta+r_I,\infty)+m(r_O,\infty)}$. Now suppose that $1-\delta_2 \geq w_l \geq 1-\delta_1$ for some arbitrarily given $\delta_2 < \delta_1$. Then we claim that, if $n$ is large, $\mathbb{E}[\tilde{C}]$ is negative. To see this note that, if $w_l$ is bounded away from 1 and $n$ is large enough, the sign of $\mathbb{E}[\tilde{C}]$ is that of the term $\frac{m(\eta+r_I,\infty)}{m(\eta+r_I,\infty)+m(r_O,\infty)} - w_l$. Thus, since choice of $\delta_1$ ensures that

$$w_l \geq 1-\delta_1 > \frac{2m(\eta+r_I,\infty)+m(r_O,\infty)}{2m(\eta+r_I,\infty)+2m(r_O,\infty)} > \frac{m(\eta+r_I,\infty)}{m(\eta+r_I,\infty)+m(r_O,\infty)}.$$

the desired conclusion follows. ■

# Appendix 2

Here we provide details regarding the link between observable matches and underlying strategy and descriptive statistics for the sample used in the microlevel analysis.

### The relationship between empirical matches and strategy

Let: $\eta$ denote the number of draws from the strategy-determined pool (either in-group or full population); $r_O > 1$ and $r_I > 1$ denote the number of draws, independent of strategy, made from the population, $N$, and the restricted in-group pool, $n_l$, respectively; $p$ denotes the probability of matching with any given draw. $\eta$, $p$, $r_O$ and $r_I$ can vary randomly across individuals but are independent of group size, population size and strategy. These parameters are not observed by the econometrician.

Let $s$ denote the strategy played by an individual where $s = 1$ for inbreed and $s = 0$ for outbreed. Let $s^\star$ denote the empirical variable where $s^\star = 1$ if all observed matches are from the in-group and $s^\star = 0$ if at least one observed match is from the out-group. We want to know: What can be learned about $s$ from the observation of $s^\star$?

$s^\star$ can be mapped to $s$ through the following relationship:

$$s^\star = (1 - \mathbb{1}[\text{outgroup match}|r_O])s + (1 - \mathbb{1}[\text{outgroup match}|\eta + r_O])(1-s)$$

where $\mathbb{1}[.|k]$ is an indicator function taking a value of 1 if the argument is true, based on $k$ draws from the full population, and 0 otherwise.

This implies that the probability of observing $s^\star = 1$ is:

$$P[s^\star = 1] = (1 - P[\text{outgroup match}|r_O])P[s = 1] + (1 - P[\text{outgroup match}|\eta + r_O])(1 - P[s = 1]).$$

Clearly, we need to derive $P[\text{outgroup match}|r_O]$ and $P[\text{outgroup match}|\eta + r_O]$ in terms of the model's parameters.

Let $n_l$ denote the (absolute) size of of the in-group for an individual of race $l$, $n_O = \sum_{l' \neq l} n_{l'}$ denote the (absolute) size of the out-group and $w_l = n_l/(n_l + n_O)$ be its relative population share.

For any draw made from the full population, the probability of drawing an out-group member is given by $1 - w_l$. The probability of any given draw being a suitable match is given by $p$. A given individual will only be observed with in-group matches if one of the the following is true:

1. For all draws from the full population, only ingroup members are drawn.
2. Of the draws from the full population that result in outgroup members, none are a suitable match.

An inbreeder makes $r_O$ independent draws (with replacement) from the full population, and an outbreeder makes $\eta + r_O$ independent draws (with replacement) from the full population. Therefore, we can write

$$\begin{aligned} 1 - P[\text{outgroup match}|r_O] &= [w_l + (1 - w_l)(1 - p)]^{r_O} \\ 1 - P[\text{outgroup match}|\eta,\, r_O] &= [w_l + (1 - w_l)(1 - p)]^{\eta + r_O} \end{aligned}$$

Letting $\zeta = w_l + (1 - w_l)(1 - p)$ we can write:

$$P[s^\star = 1|w_l,\, n_l] = P[s = 1|w_l,\, n_l]\zeta^{r_O} + (1 - P[s = 1|w_l,\, n_l])\zeta^{\eta + r_O}$$

Taking the partial derivative of $P[s^\star = 1|w_l,\, n_l]$, conditional on $w_l$, we can say the following:

$$\frac{\partial P[s^\star = 1|w_l,\, n_I]}{\partial n_l} = \frac{\partial P[s = 1|w_l,\, n_l]}{\partial n_l}(\zeta^{r_O} - \zeta^{\eta + r_O})$$

The object in parenthesis on the right-hand-side is strictly positive. It follows that the sign of $\frac{\partial P[s^\star = 1|w_I,\, n_I]}{\partial n_I}$ is determined by the sign of $\frac{\partial P[s = 1|w_I,\, n_I]}{\partial n_I}$. However, the magnitude will be an underestimate.

A similar inference cannot be drawn from the behavior of $\frac{\partial P[s^\star = 1|w_I,\, n_I]}{\partial w_I}$. To see this write:

$$\begin{aligned} \frac{\partial P[s^\star = 1|w_I,\, n_I]}{\partial w_I} &= \frac{\partial P[s = 1|w_I,\, n_I]}{\partial w_I}([\zeta^{r_O} - \zeta^{\eta + r_O}) \\ &+ (\eta + r_O)p\zeta^{\eta + r_O - 1} + P[s = 1|w_I,\, n_I]p(r_O\zeta^{r_O - 1} - (\eta + r_O)\zeta^{\eta + r_O - 1}) \end{aligned}$$

The second right-hand-side term is positive, but the remaining terms are indeterminate. $\frac{\partial P[s = 1|w_I,\, n_I]}{\partial w_I}$ cannot be signed based on $\frac{\partial P[s^\star = 1|w_I,\, n_I]}{\partial w_I}$.

## Descriptive statistics

Table 4: Racial matches: Add Health friendship nominations and U.S. population census marriages

| *Friendship nominations* | | | | | | |
|---|---|---|---|---|---|---|
| | | | Race of nominee | | | |
| Race of nominator | White | Black | Hispanic | Asian | Other | |
| White | 81.41 | 2.03 | 6.11 | 1.64 | 7.96 | |
| Black | 7.17 | 72.63 | 9.48 | 1.00 | 8.38 | |
| Hispanic | 25.72 | 9.96 | 52.28 | 3.68 | 7.40 | |
| Asian | 24.15 | 3.81 | 11.38 | 50.00 | 9.91 | |
| *Marriages* | | | | | | |
| | | | Race of husband | | | |
| Race of Wife | White | Black | Native | Asian | Hispanic | Other |
| White | 93.69 | 1.06 | 0.43 | 0.73 | 2.58 | 1.51 |
| Black | 1.64 | 96.66 | 0.10 | 0.16 | 0.73 | 0.71 |
| Native | 53.40 | 6.23 | 25.80 | 1.81 | 6.65 | 6.11 |
| Asian | 28.32 | 2.29 | 0.32 | 63.10 | 2.63 | 3.34 |
| Hispanic | 19.89 | 2.21 | 0.28 | 0.76 | 73.97 | 2.89 |

Each cell reports the percent of total nominations (marriages) for nominee (husband) race by nominator (wife) race.

Table 5: Descriptive statistics: Add Health friendship nominations and U.S. population census marriages

| | Mean | Std. Dev. | Max. | Min. |
|---|---|---|---|---|
| *Friendship nominations* | | | | |
| Students (per school) | 1 178 | 552.54 | 20 | 2 284 |
| Sex = female | 0.52 | 0.51 | 0 | 1 |
| Race = White | 0.62 | 0.48 | 0 | 1 |
| Race = Black | 0.16 | 0.37 | 0 | 1 |
| Race = Hispanic | 0.16 | 0.37 | 0 | 1 |
| Race = Asian | 0.05 | 0.22 | 0 | 1 |
| Grade | 9.55 | 1.73 | 0 | 12 |
| Nominations (per student, total) | 7.81 | 2.57 | 1 | 10 |
| Nominations (per student, observed) | 5.20 | 2.55 | 1 | 10 |
| Out-group nominations per student | 0.87 | 1.46 | 0 | 10 |
| Schools (total) | 78 | | | |
| Students (total) | 55 676 | | | |
| *Marriages* | | | | |
| Race = White | 0.72 | 0.45 | 0 | 1 |
| Race = Black | 0.17 | 0.37 | 0 | 1 |
| Race = American Indian | 0.01 | 0.08 | 0 | 1 |
| Race = Hispanic | 0.02 | 0.14 | 0 | 1 |
| Race = Asian | 0.09 | 0.29 | 0 | 1 |
| Educ. < High school | 0.11 | 0.31 | 0 | 1 |
| Educ. = High school | 0.35 | 0.48 | 0 | 1 |
| Educ. > High school | 0.54 | 0.50 | 0 | 1 |
| Age | 33.86 | 7.36 | 20 | 49 |
| City population | 354 538 | 721 988 | 80 800 | 8 184 900 |
| City-year observations | 620 | | | |
| Marriages (total) | 528 489 | | | |