A Simple Model of Homophily in Social Networks

Sergio Currarini* Fernando Vega Redondo†

July 11, 2011

Abstract

We study the formation of social ties in a context where agents are heterogeneous and take actions that shape the size and the type-composition of their meeting pools. Our aim is to provide simple microfoundations to the important phenomenon of homophily, i.e. the tendency of agents to connect with others of the same type. In particular, we focus on the incentives and the forces that lead agents to distort their meeting rates away from the type composition at the population level. The equilibrium behavior implied by our model is of a threshold type, with agents "inbreeding" (i.e. mostly meeting their own type) iff their group is above certain size. This yields a pattern of in-group and cross-group ties that are consistent with empirical evidence from two diverse instances of social networks such as high school friendships and interethnic marriages.

Keywords: Homophily, social networks, segregation.
JEL Classification: D7, D71, D85, Z13.

*Dipartimento di Economia, Università di Venezia. Email: s.currarini@unive.it.

†European University Institute (Florence) and Instituto Valenciano de Investigaciones Económicas. Email: Fernando.Vega@eui.eu.

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. The presence of homophily has important implications on how information flows along the social network (see e.g. 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.

Naturally, one basic factor affecting the distribution of in- and cross-group links must be how agents are spread across different (largely homogenous) groups. For, as pointed out by Blau (1977), the relative sizes of groups affect the meeting opportunities of agents.[1] If ties were formed in an unbiased manner and meeting was uniform, agents would end up meeting a fraction of members of a given group reflecting the group's population share. So, larger groups, whose members are met with higher probabilities, would tend to make a smaller fraction of ties with dissimilar agents. This is a "baseline" form of homophily that affects the distribution of in-group ties even if neither the agents' attitudes nor the meeting mechanism distorts at all the formation of ties.

What is interesting about the empirical evidence is that many social networks display homophily in excess of this baseline level. This suggests that the generative process of tie formation is not properly described by uniform assortment. Rather, it is the result of significant biases, either in agents' preferences and/or in agents' meeting opportunities. The presence of such biases has in turn important implications for policy. For example, it implies that despite efforts to bring about a balanced type distribution in the population (as, for example, it has been done in schools), pronounced segregation patterns may well persist due to the process through which agents establish their social ties. Thus, for the successful implementation of a policy that aims at, say, inter-ethnic integration, it is crucial to have a proper understanding of such endogenous forces at work.[2]

Homophily has been the object of study in the sociological literature, at least since the work of Lazarsfeld and Merton (1954) – see e.g. Marsden (1987, 1988), Moody (2001), or the comprehensive survey in McPearson, Smith-Lovin and Cook (2001). This literature has been of a mostly empirical character. Recently, however, there have been several contributions in economics that have incorporated a theoretical dimension to the analysis. Among these, the closest to our concerns is Currarini, Jackson and Pin (2009) – henceforth referred to as CJP – which proposes a model

[1] The idea that a rise in the encounter opportunities enhances the probability of tie formation is already present in Alport's (1954) contact theory

[2] Our model suggests, for example, that integration will tend to be low if there are a number of relatively large groups in any given population. The integration of minorities, therefore, would be better served by type dispersion rather than by efforts to bring about concentration on a few large groups.

of tie formation that is tested against data on friendship networks obtained at racially heterogeneous American high schools. Its primary concern is to disentangle the incidence of preference and meeting biases in the formation of social networks. One of its key insights is that the patterns of homophily observed in friendship networks *cannot* be explained alone by homophilous *preferences*, i.e. a desire to connect to others of the same type. In their setup, therefore, even though preference bias does play an important effect (for example, in affecting the number of friends per capita), it is only if individuals display some *meeting* bias that the observed pervasive homophily can in the end materialize.

The previous discussion suggests that a satisfactory model of homophily should include a suitable account – preferably grounded on basic and simple principles – of the forces that lead to meeting biases in social systems. This paper is a preliminary attempt in this direction. We provide a microfounded model where such meeting biases derive from a very simple mechanism of tie formation whose key ingredients are as follows.

- Agents derive positive utility from the number of *distinct* ties they enjoy.
- In order to form these ties, they can rely on a fixed number of meeting draws obtained from a certain pool of agents.
- Their single decision is to select the pool on which to conduct their quest for ties. Two options are available:
  (a) they restrict to individuals like their own – a decision we call *inbreeding*, or
  (b) they extend their efforts to the whole population, which we label *outbreeding* and entails a fixed cost.

The dichotomous dilemma given by (a)-(b) embodies conflicting incentives. On the one hand, there is the cost associated to outbreeding, which may reflect cultural, geographical, or linguistic barriers to accessing other types. And, on the other hand, inbreeding has the drawback that it limits the effectiveness of agents' efforts, by increasing the chance of redundant draws.

The previous considerations generate a tension between inbreeding and outbreeding, which turns out to be resolved in a different manner in large and small groups. Specifically, our analysis shows that there exists a threshold for group size such that groups above it inbreed while those below outbreed. Thus a key insight that results from our simple model is that the meeting bias arising at equilibrium is not uniform: it applies just to large groups. Only their members, that is, find it optimal not to pay the cost of outbreeding and, consequently, end up meeting agents of their own type alone.

Our stylized theoretical framework captures an essential trade-off that should play an important

role in general processes of tie formation among heterogeneous agents.[3] The general formulation can be specialized in a number of different directions, to accommodate for different specific features of the meeting mechanism and the induced payoffs. For concreteness, we shall focus on two polar cases: a one-sided drawing scenario with one-sided payoff flows, and another one where both drawing and payoff flows are two sided. In the one-sided scenario, outbreeding agents access the population at large (outbreeding or not), and positive payoffs are obtained only by those who *actively* form the links. This is representative of situations in which agents can unilaterally decide to establish a profitable connection, as for some instances of information acquisition (e.g. internet browsing). By contrast, in the two-sided setup, outbreeders only meet outbreeders. This represents situations where mutual consent or some coordinated action is needed in order to form a link. Natural instances of such a two-sided mechanism occur, for instance, when interaction requires that agents move to a common physical location (say "downtown"), or learn a common *lingua franca*. The specific results are somewhat different in each scenario but, as we shall explain, they all display the same essential implications.

To test empirically our model, we take it to data gathered in two quite diverse contexts: high school friendships from the widely used AddHealthdataset, and inter-ethnic marriages obtained from U.S. census data.[4] For the first context, CJP found that the so-called Coleman Index (a normalized measure of homophily) depends nonlinearly on group size, with maximal values for middle-sized groups and low positive values for small and large groups. In this paper, we report a similar pattern for U.S. marriages.[5] This suggests that such a non monotonic dependence of homophily on size may be a common feature to be expected in a wide range of social contexts. It is therefore reassuring to confirm that our very basic model is capable of reproducing it.

The equilibrium analysis of our model also yields a good number of additional empirical predictions that are consistent with the aforementioned data sets on friendships and marriage. Here, for the sake of focus, we highlight only three of the predictions the model will be seen to deliver.[6]

- A first one concerns the share of in-group ties as a function of the absolute size of groups.

[3]This trade-off arises starkly in our simple context but should also be present in other more elaborate setups. By way of illustration, consider a dynamic procedure whereby new ties are formed over time by relying on the intermediation of existing ties. In this context, the network of existing ties in small groups will typically display high clustering, in turn yielding the sort of detrimental redundancy induced by our basic model.

[4]See Subsection 4.2 for a detailed description of these datasets.

[5]See also Rogers and Bramoullé (2009).

[6]The model also predicts, for example, that the number of effectively formed ties grows with group size for small groups but the effect becomes insignificant for large groups. This is in line with our empirical evidence, but stands partly in contrast with the model by CJP that predicts a positive effect of group size throughout. Another interesting prediction of the model that derives only from its two-sided variant is that cross-group ties are to be primarily found between agents of small groups. This conclusion is well aligned with our evidence on marriages but only weakly so with that on friendships, which suggests that the two-sided formulation may be more adequate for the former case but less so for the latter.

The prediction of the model is that this share should exhibit an abrupt upward jump at some threshold level for group size, beyond which it should remain at a uniformly high level.

- A second prediction concerns the difference between the share of in-group ties and the population share of inbreeding groups. The model predicts that this difference should be linearly decreasing in the population share.

- A third prediction of our model concerns the role of absolute population size on the onset of homophilous behavior. Conditional on the relative shares of each group, the prediction is that the larger is the overall population the (weakly) higher is the number of inbreeding groups.

The above implications of the model illustrate its potential to shed light on the important phenomenon of homophily in social networks. Admittedly, the theoretical framework is abstract and "minimalist." This, however, serves to underscore what, in different specific forms, may be a principle at work in a wide range of different social contexts. In essence, it is the idea that one's own group defines the boundaries of costless (or cheap) interaction, but the range of such an interaction may be limited if the group is small. This, in turn, may justify incurring the cost of external "outbreeding," which indeed opens up the range of interaction but is costly. In a nutshell, the main contribution of the paper is to show that this very basic trade-off may bring us quite far in understanding the pattern of homophilous behavior observed in many social contexts.

The remainder of the paper is organized as follows. Section 2 describes the model, including the strategies and payoffs defining the underlying meeting game. Section 3 characterizes the equilibrium behavior of agents, as a function of the size of their respective groups. In Section 4, the core of the paper, we apply the model to shed light on the phenomenon of homophily, organizing the discussion in three parts. First, we describe different measures used to quantify the degree of homophily in a population. Second, we describe our evidence on high-school friendships and marriages. Third, we discuss the match between this evidence and our theoretical predictions. The main analysis of the paper is complemented in Section 5 with some further discussion on the contrasting implications of one- and two-sided models and the role of population and group size. Section 6 concludes the main body of paper. For the sake of exposition, all proofs are included in an Appendix.

2 The Model

We 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.), that we call "type". Groups are indexed by l, and we denote by n_l the size of group $l = 1, 2, ..., q$. Each agent i 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 ν $(\leq \eta)$ 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). We will refer to the first type of activity as "inbreeding", and to 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., traveling, learning a language, or changing one's habits)

The inbreeding/outbreeding decisions taken by all agents constitute their strategies in the game. They determine the meeting pool each of them accesses, which in turn shapes the probability distribution over *distinct* partners they face, and thus their expected payoffs. In what follows, we define the game formally by presenting in turn each of its aforementioned components.

2.1 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 $\mathbf{s} \equiv (s_i)_{i \in N} \in \{I, O\}$ 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 η independent draws with replacement, when the *size* of the pool is $\theta_i(\mathbf{s})$.

As advanced, we shall distinguish two different scenarios concerning how the meeting pool of an agent is shaped by the strategy profile $\mathbf{s}$. Each scenario is captured by corresponding specifications of the functions $\theta_i(\cdot)$ and $\tilde{\nu}(\cdot)$ shaping, respectively, the size of the meeting pool and the meeting opportunities.

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, for example, an agent outbreeds by taking the initiative and visiting the "locations" (geographic or virtual) where people from other groups live. 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 meeting pool accessed in this context by that agent given the strategy profile $\mathbf{s}$. Then we posit:[7]

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

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

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

In line with the postulated one-sidedness of the meeting mechanism in this case, the payoff flows will be assumed to be one-way as well.[8] 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 η tries when the pool size is θ.

In contrast, a *two-sided* context is one where the meeting pool of any outbreeding group consists of those groups that have themselves chosen to outbreed. This gives rise to an alternative function $\overleftrightarrow{\theta_i}(\mathbf{s})$ that specifies 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 posit:

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

Such a two-sided scenario models situations in which, for example, all outbreeders move to some fixed location ("downtown") where only outbreeders meet, or they switch to a common *lingua franca* distinct from the one each of them originally speaks. Under these conditions, 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 $\tilde{\nu}_{\leftrightarrow}(\eta, \theta)$ used to capture this situation must be different from before. It captures the random number of distinct meetings obtained in a pool of θ agents when

(a) each agent makes η independent draws with replacement, and

(b) a meeting is said to occur between two agents, i and j, when either a draw by i selects j or *viceversa*.

A key aspect of both the one-sided and two-sided scenarios is that a larger pool brings about richer meeting possibilities, in the sense that the range of distinct partners one can meet is likely to be wider. This feature introduces the basic tradeoff between the inbreeding and outbreeding decisions that is at the core of our model. Mathematically, such richer possibilities are captured by an appropriate ranking of the two families of random variables, $\{\tilde{\nu}_{\rightarrow}(\eta, \theta)\}_{\theta \in \mathbb{N}}$ and $\{\tilde{\nu}_{\leftrightarrow}(\eta, \theta)\}_{\theta \in \mathbb{N}}$, as parametrized by the size of the meeting pool θ. Indeed, as we shall explain in Section 4, 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.

[8] 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.

- In the two-sided scenario, larger meeting pools yield a higher expected number of distinct draws.

Of course, the second criterion is much weaker than the first. But, as we shall see, both induce, under suitable assumptions on preferences (naturally, stronger in the second case), an interestingly sharp tradeoff between the inbreeding and outbreeding options.

REMARK **1** *As it turns out, the only feature of our alternative meeting scenarios that is key to establishing all our results is that, as the pool of interaction possibilities grows, so does the entailed payoff potential. It follows, therefore, that any of the alternative specifications that satisfy such an intuitive requirement will yield an equivalent analysis of the problem.*

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. Then if, as it is natural to postulate, payoff-enhancing diversity grows as the pool of alternative partners expands, the desired effect of pool size follows. As a second possibility, it could be assumed that new partners are found through existing partners in a social network whose size depends (as in our model) on the breeding decisions taken. Then, the effectiveness of searching for new partners must depend (negatively) on network clustering, which in turn is well know to decrease with size if the network can be suitably described as a a complex and largely random network (see e.g. Vega-Redondo (1997).

2.2 Preferences

Now we describe agents' preferences upon their meeting outcomes. Denote the number of distinct meetings enjoyed by any given agent i by ν_i. Given that each distinct partner is 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 Neuman-Morgenstern (vNM) utility $U : \mathbb{N} \cup \{0\} \to \mathbb{R}$, where we normalize $U(0) = 0$ and posit that

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

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

Thus we only demand that no suitable partner is worse than one such partner, but there may well be saturation beyond that point. Thus, in general, the theoretical framework may accommodate different applications, say 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 U may be postulated to level at one. [9]

[9]Note here that the type composition of one's meetings is inessential for welfare, 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, and introducing further trade-offs that can be of interest in future research.

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

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

But, even given the pool size θ faced by agent i, the number ν_i of her distinct partners is uncertain from an *ex ante* viewpoint. It is given by the random variable $\tilde{\nu}(\eta, \theta)$, 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 θ. This gives rise to the expected utility $W(\theta)$ for a typical agent i when she faces a meeting pool of size η defined as follows:

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

where $P_{\eta,\theta}$ denotes the probability distribution associated with the random variable $\tilde{\nu}(\eta, \theta)$.

2.3 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.[10]

The payoff of the agent is defined as follows:

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

where $P_{\theta_i(\mathbf{s})}(\cdot)$ stands for the probabilities over the number of distinct meetings that is induced by the random variable $\tilde{\nu}(\theta_i(\mathbf{s}))$, $c(s_i) = c > 0$ if $s_i = O$ and $c(s_i) = 0$ if $s_i = I$, and recalling that p is the probability that any given partner be suitable, while $U(\cdot)$ defines the vNM utility of agents over the number of suitable partners.

For simplicity, our equilibrium analysis will focus throughout on profiles s that are group-symmetric, i.e. where $s_i = s_j$ whenever i and j belong to the same group l. Within this class, the population behavior can be fully described by the q-tuple $\gamma \equiv (\gamma_1, \gamma_2, ..., \gamma_q)$ that specifies the common choice $\gamma_l \in \{I, O\}$ for every agent in any of the groups $l = 1, 2, ..., q$. The induced meeting pools are denoted by $\{\theta(\gamma)\}_{l=1,2,...,q}$.

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

The payoff of any typical agent i of group l is now given by:

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

3 Equilibrium

We now characterize the group-symmetric Nash equilibria of our game. Our key conclusion in this respect is that the equilibrium behavior of a group is fully dependent on whether it is large or small, as compared to a certain threshold. More specifically, we find that all groups whose size is smaller than the equilibrium 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 stating the result that applies to 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. 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, that is, we shall prove the following auxiliary lemma.

LEMMA **1** *For any given θ, θ', 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 τ^* 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 τ_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: the benefit of outbreeding now depends on the *endogenous* size of the outbreeders' pool. This, in contrast with the one-sided scenario, leads to 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 their own type alone. Despite the possibility of such an "deadlock," the following proposition establishes that as long as the small groups make up for a large enough (that is, non negligible) share of the whole population, the existence of a positive-threshold equilibrium carries over to this case as well.

THEOREM **2 (Threshold Equilibrium – two-sided scenario)** *Consider the two-sided scenario, and assume that the utility function U is linear. 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, if the outbreeding cost satisfies $c < V(\eta)$, for any $\alpha > 0$ there exists some $\hat{n}$ such that if $n \geq \hat{n}$ and $\sum_{l:n_l<\tau*} n_l > \alpha n$, then a threshold equilibrium exists with $\tilde{\tau} = \tau^*$, where τ^* is the threshold defining the (unique) equilibrium in the one-sided scenario.

The two-sided scenario opens up, in contrast with the one-sided case, the possibility of miscoordination equilibria. In an intuitive sense sense, all equilibria associated to thresholds $\tau < \tau^*$ embody a certain manifestation of coordination failure, the extreme version of it being the aforementioned trivial full-inbreeding equilibrium induced by $\tau = 0$. Again in this case, the result builds upon the fact that 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 θ, θ', 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 the two sided scenario, the size of the pool where search occurs affects the (random) number of potential matches in two distinct ways. First, the redundancies that induce wasteful meetings tend to weaken as the pool gets larger, as in the one-sided model. Second, a larger pool implies a smaller probability of being found by each of the other agents in the pool. This second effect introduces an additional source of complexity in the problem, and forces us in Lemma 2 to weaken

(in comparison with Lemma 1) the criterion used to rank the meeting distributions.[11] This, in turn, also explains the stronger assumption on the utility function made in Theorem 2 where, in addition to the basic assumption of monotonicity, we require that the utility function be linear in the number of distinct meetings. Throughout the paper, we shall assume that in a two-sided meeting scenario the conditions contemplated in Theorem 2 are met and the population plays some threshold equilibrium in which the number of outbreeding agents is a nonvanishing fraction of the whole population.

4 Homophily: theory and evidence

As we discuss in this section, the equilibrium inbreeding/outbreeding choice derived from the model has strong empirical implications, which are then compared with empirical evidence. Our discussion is structured as follows. First, in Subsection 4.1, we introduce the measures that capture some of key features of the matching pattern, both within and across groups. Second, in Subsection 4.2, we briefly introduce the sources of empirical evidence (on friendship and marriages) that will be used to test our model. Third, in Subsection 4.3 we derive some of the main theoretical implications of the model and contrast them with our data.

4.1 Measuring homophily

To start with, a basic measure of the degree of homophily of any given group l can be obtained as follows. 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. Then, the ratio $\frac{m_{ll'}}{m_l}$ measures the representation of type l' matches in the total matches of group l, with the particular case where $l' = l$ giving rise to what is called the *Homophily Index* of group l, $H_l \equiv \frac{m_{ll}}{m_l}$.

Next, we are interested in comparing the homophily index of any group l with the proportion of own-group matches that would result if there were no "bias" and that proportion were equal to the population share of group l. Such a comparison is simply captured by what we shall call *excess homophily*, which is defined as the difference $H_l - w_l > 0$ for each group l.[12] If prevailing ties where generated by a uniform random assortment, we would expect $H_l = w_l$ and therefore a zero excess homophily for all types .

When it comes to comparing the homophily of different groups, however, the simple difference $H_l - w_l$ would provide a distorted picture of groups' attitudes to inbreed. For, in effect, groups with very large size w_l could never experience large excess homophily due to the simple reason

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

[12] The 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 the "inbreeding" choice of agents in our model.

that its maximal potential value, $1 - w_l$, is small to begin with. The index proposed by Coleman (1958), and recently employed in various papers (see Currarini, Jackson and Pin (2009), Bramoullé and Rogers (2009)) addresses the problem by normalizing the excess homophily of group l by its maximal value $1 - w_l$:

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

This will be called the *Coleman (Homophily) Index* and will attract much of our attention in what follows.

4.2 Friendships and marriages

Our theory will be brought to the data for two types of social networks: high school friendships and marriage, for which race and ethnicity are very significant dimensions. Our aim will be to assess to what extent the threshold structure of equilibrium predicted by the model can explain some of the key empirical regularities of social ties both within groups (in-group ties) and between groups (cross-group ties).

For high school friendships, we consider the national sample of American high schools covered by the Add Health dataset, which has been extensively studied in many sociological works on homophily (see, for instance, Moody (2001)), and more recently by Currarini, Jackson and Pin (2009, 2010) in their economic model of friendship.[13]. This dataset reports friendships nominations made by students, in order of importance and by gender. It can be used, therefore, to reconstruct the full network of friendships, allowing as well to keep track of various individual characteristics such as race,[14] income, gender, and various other behavioral traits. An observation refers here to a given ethnic group in a given school of the sample.

On the other hand, our study of interethnic marriages is based on the database IPUMS (Integrated Public Use Microdata Series), which records personal census data for the U.S. from 1850.[15] An observation in this dataset identifies a triple " ethnicity, year, geographical area" (for example: Indians, in 1980, in the New York Urban State). We cover the years 1960-2000, with 10 years intervals, and years 2000-2007 on a yearly basis. Originally, each state is identified with a separate

[13]The National Longitudinal Study of Adolescent Health (AddHealth) is a longitudinal study of a nationally representative sample of adolescents in grades 7–12 in the United States during the 1994–95 school year. Data files are available from Add Health, Carolina Population Center (addhealth@unc.edu).

[14]Racial groups are Whites, Blacks Hispanic and Asians.

[15]IPUMS consists of a series of compatible-format individual-level representative samples of the American population (one per cent of it) for the years 1850-1880, 1900-2000, together with the American Community Surveys of 2000-2007, and the Puerto Rican Community Surveys of 2005-2007. It is produced and distributed by the Minnesota Population Center. Please quote the dataset as follows: "Steven Ruggles, Matthew Sobek, Trent Alexander, Catherine A. Fitch, Ronald Goeken, Patricia Kelly Hall, Miriam King, and Chad Ronnander. Integrated Public Use Microdata Series: Version 4.0 [Machine-readable database]. Minneapolis, MN: Minnesota Population Center [producer and distributor], 2008."

"marriage market." However, if a state has a city with more than 500.000 people, we split it in two: urban and rural, under the hypothesis that each shows different patterns. Overall, eight ethnicities are considered: White, Hispanic, Black, Native, Chinese, Japanese, Indian, Other Asian.

4.3 Patterns of homophily

Here we present and discuss the empirical evidence (both for friendships and marriages) on each of the measures of homophily described in Subsection 4.1. Thus we address in turn the Homophily Index, the Excess Homophily, and the Coleman Index for each of our two empirical contexts and discuss whether the evidence can be understood in terms of the theoretical implications of our model.

4.3.1 Homophily Index

Empirical Evidence

As mentioned, we start with the basic index of homophily measured by the fraction of the total number of ties of this group that are associated to individuals of that same group. We report in Figure 4.3.1 the empirical pattern of this index for both friendships (left) and marriages (right) with respect to groups' sizes. In these diagrams, as well as those to follow, each dot refers to a different observation in the relevant dataset. Thus, for friendships, each dot corresponds to a certain ethnic group in a particular school; for marriages, each dot refers to a given group, region, and year.

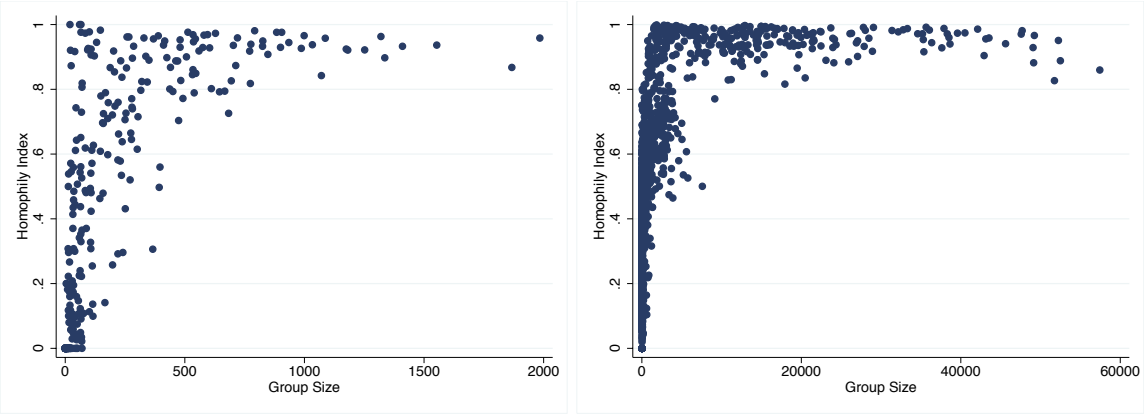

Figure 1: Homophily Index: High School Friendships (left) and Marriages (right).

For the case of friendships we can observe a sharp structural change in the relationship at an approximate group size of 300 students (for a maximum size of around 2000) while for marriages such a change occurs at an approximate group size of 4000 couples (for a maximum size close to 60000). To obtain a statistical assessment of such structural changes, we include intercept and

slope dummies in the regressions of the homophily index against group size, which take value 1 for group size larger than 300 and 4000, respectively. All coefficients and dummies are statistically significant[16]. In particular: for friendships, the constant term shifts from 0.09149 to 0.7987, while the slope decreases from 0.003 to 0.0001; for marriages, the constant term shifts from 0.303 to 0.893, while the intercept decreases from 0.0002 to 0.000002.

Theory

The structural break in the homophily index displayed by Figure 4.3.1 is in line with the threshold behavior established by Theorems 1-2. The theory implies that the expected fraction of distinct meetings (and thus suitable partners) that belong to one's own group should experience an abrupt increase when the size of the group exceeds a certain threshold. Operationally, therefore, if we establish a correspondence between "suitable partners" an "actual (or observed) matchings," the structural break found in the empirical evidence reflects a similar phenomenon.

The evidence, however, shows significant dispersion in the homophily associated to each group size, and this raises the question of what may underlie such variability. One reason for it stems from the fact that our theoretical analysis only applies to asymptotically large populations. And, of course, in the empirical scenarios considered, not only the groups but also the overall population are finite, displaying as well relatively small sizes in some cases. Thus, in general, we must expect that the variables of interest should deviate from the expected values, these deviations being wider and more frequent the smaller the group size and the smaller the overall population. This may partially explain why the dispersion found for friendship in high schools (where group and population sizes are relatively small) is narrower than for marriages (where they are much larger).

But there are, of course, many other reasons why one should expect some variability in the homophily index displayed by groups of similar size. Among these, we may point to any unmodeled friction that limits the ability of agents to implement a pure breeding strategy – e.g. some bias in the spatial distribution of agents.[17] Rather than attempting to model such a friction in any detail, we shall aim at capturing it in an abstract reduced form. Specifically, we shall simply 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.

The above postulate ensures that some amount of inbreeding as well as outbreeding occurs for all groups, independently of their size and intended behavior. Since (F) is conceived as a "perturbation," the numbers r_I and r_O are to be thought as small (> 1). This notwithstanding, (F) can be seen to have several interesting implications, as we now explain.

[16] When not otherwise specified, we mean a 99% significance level.

[17] By way of illustration, suppose that agents of different types live in separate areas. Then, it will be unavoidable that they meet a significant fraction of their own type, even if they have decided to outbreed. Or, reciprocally, even if all groups inbreed it will be unavoidable that there is some contact across them if they have to share some common resource or jointly participate in production activities.

First, it readily implies that, in expected terms, the homophily index of any "small" group grows with its size. For the sake of completeness, this is stated in the following result.

PROPOSITION **1** *Consider any two groups l and l' of given (finite) sizes, $n_l < n_{l'}$. There exists some $\hat{n}$ such that if $n \geq \hat{n}$, then $\mathbb{E}[\tilde{H}_{l'}] > \mathbb{E}[\tilde{H}_l]$, where $\tilde{H}_l$ and $\tilde{H}_{l'}$ are the random variables specifying the homophily indices of groups l and l'.*

A second implication of (P) is that, *ex ante*, the homophily index of a small inbreeding group is given by a random variable whose probability distribution displays a wide support if η is large.

PROPOSITION **2** *Given any (finite) χ, consider any inbreeding group l with its size n_l satisfying $\chi \geq n_l > r_I + \eta$. There exist some $\hat{n}$ such that if $n \geq \hat{n}$, then the probability distribution of the random variable $\tilde{H}_l$ has support included the set $B \equiv \{\frac{1}{r_O+1}, \ldots, \frac{r_I+\eta}{r_I+r_O+\eta}\}$ with $\Pr_{\tilde{H}_l}[x] \geq \delta$ for all $x \in B$ and some $\delta > 0$.*

The previous two results help explain two of the features highlighted in connection with Figure 4.3.1. The first one, Proposition 1, addresses the positive slope of the lines obtained when regressing the Homophily Index on group size, both before and after the structural break. On the other hand, Proposition 2 provides a basis to understand why, for low group sizes around the structural break, we find a wide dispersion in the homophily index. Note, however, that as the group size gets large enough our theory predicts that the dispersion of the homophily index shrinks around a value of close to one, as indeed observed in Figure 4.3.1.

4.3.2 Excess Homophily

Empirical Evidence Figure 2 depicts the *excess homophily* of groups as a function of their respective population frequencies, for both friendships (left panel) and marriages (right panel). This figure highlights the following two-part pattern:

(a) the excess homophily is increasing (and steeply so) for groups covering less than 10% of total population;

(b) beyond that threshold, it decreases *linearly* with an absolute slope less than one.

Again, as explained above, this is particularly clear in the case of marriages where the larger sizes involved allow for a sharper manifestation of the phenomenon.

Theory To account for Part (a) of the aforementioned pattern, we can rely on considerations analogous to those underlying Proposition 1 to establish the following result.

PROPOSITION **3** *Consider any two outbreeding groups l and l' with $n_l < n_{l'}$. There exist some $\hat{n}$ such that if $n \geq \hat{n}$, then $\mathbb{E}[H_{l'}] - w_{l'} > \mathbb{E}[H_l] - w_l$.*

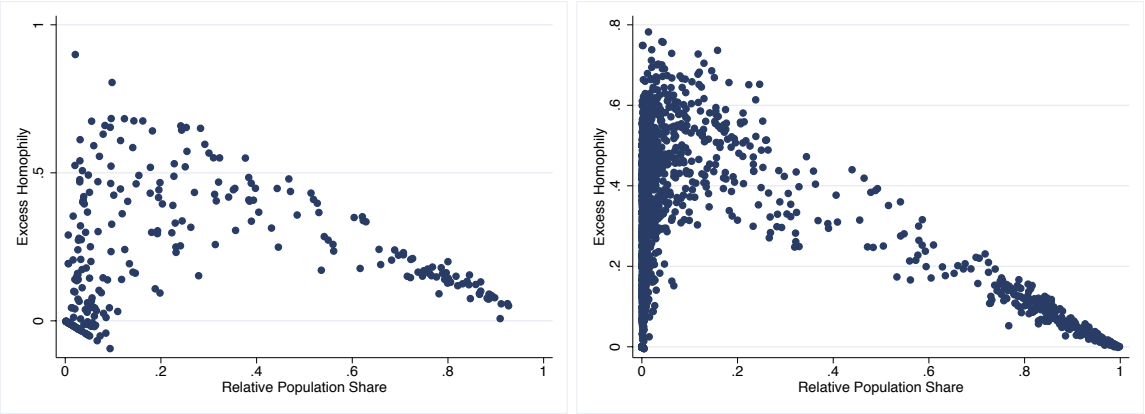

Figure 2: Excess Homophily by Relative Group size: High School Friendships (left) and Marriages (right).

As for its Part (b), the gist of it is largely captured in our theoretical framework by the following proposition.

PROPOSITION **4** *Consider any two inbreeding groups l and l' with $w_{l'} > w_{l'} \geq \vartheta > 0$. There exist some $\hat{n}$ and some α with $0 < \alpha < 1$ such that if $n \geq \hat{n}$, then $\mathbb{E}[H_{l'}] - \mathbb{E}[H_l] = -\alpha[w_{l'} - w_l] + \delta(n)$, where $\delta(n) \to 0$ as $n \to \infty$.*

Thus we find that, at a qualitative level, the model delivers predictions concerning the dependence of excess homophily on group shares that are quite in line with our empirical evidence

4.3.3 Coleman Index

Empirical Evidence Additional insights can be gained by considering the Coleman Index, as was formally defined in (8). Figure 3 depicts the behavior of this index in our empirical evidence for both high school friendships (left panel) and marriages (right panel), associating each value of the index of a particular group to its relative size in the corresponding population. We fit lines for each case by regressing the index C_l on population shares w_l and the square of these shares. [18] The non linear and non monotonic pattern found in the left panel was first identified for friendships in CJP; Figure 3 shows that this trend also characterizes U.S. marriages. In particular, the Coleman index takes maximal values for middles sized groups and lower values for very small and very large

[18] Details of the regressions are as follows (t-statistics are in brackets): for friendships: $C = .4 + \underset{(16)}{2.1} w - \underset{(-15)}{2.2}\, w^2$; for marriages: $C = 0.29 + \underset{(25.4)}{0.76}\, w - \underset{(-22.3)}{0.70}\, w^2$. In both regressions, both coefficients for w and w^2 are statistically significant at 99% level; the constant term is significant only for marriages.

groups.[19]

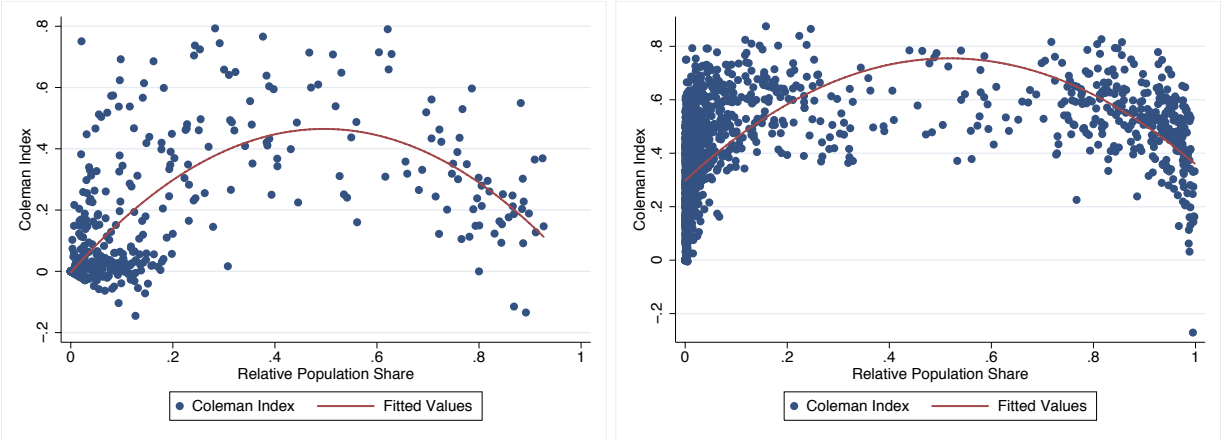

Figure 3: Coleman Homophily Index: High School Friendships (left) and Marriages (right).

Theory While the increasing branch of the parabola is an immediate consequence of the increasing pattern of excess homophily (see the discussion in the previous sections and the proof of Proposition 6 below), the decreasing part requires some additional analysis.

The following propositions will show that, if frictions are small compared to total meeting intensity (and the total population is large), the threshold equilibrium behavior induced by our model qualitatively generates the full non-monotonic pattern observed for the Coleman Index in our empirical evidence. Succinctly, they establish that (i) this index starts at very low levels, (ii) rises with size for very small groups, (iii) becomes close to 1 for intermediate-sized groups, and (iv) thereafter decreases with size, finally becoming very small (and negative) as relative size approaches 1.

For notational convenience, let us use the general notation $m(x,\theta) \equiv \mathbb{E}[\tilde{\nu}(x,\theta)]$ to denote the expected number of distinct meetings obtained when the number of draws is x and the pool size is θ.[20] Furthermore, we shall denote by $\tilde{C}_l$ the ex-ante random variable induced by our model for any given type l (a function of the corresponding threshold equilibrium in place and the group sizes). Our first result here asserts that very small groups (thus outbreeding ones), display a very small Coleman Index if frictions are small – in particular, r_I is small relative to η.

PROPOSITION **5** *Consider any outbreeding group l of given size n_l. There exists some $\hat{n}$ such that if $n \geq \hat{n}$, then the support of $\tilde{C}_l$ is bounded above by $\frac{r_I}{\eta}$.*

[19]The main difference between the right and the left panels of figure 3 is that in the one to the right 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.

[20]Recall that, in expected terms, suitable partners are proportional to distinct ones.

The next result still pertains to outbreeding groups and shows that, for these groups, the expected Coleman Index is monotonically increasing in size.

PROPOSITION **6** *There exist some $\hat{n}$ such that if $n \geq \hat{n}$, any two outbreeding groups l and l' with $n_l < n_{l'}$ satisfy $\mathbb{E}[C_l] < \mathbb{E}[C_{l'}]$.*

The next result concerns groups of "intermediate size" that inbreed. By this we mean those groups that are so large that they do *not* find it worthwhile to pay the outbreeding cost c but still represent a relatively small fraction of the whole population. For these groups, as we next state formally, the expected Coleman Index is arbitrarily high if meeting frictions are small.

PROPOSITION **7** *Given any $\epsilon > 0$, there exist some positive δ_1, δ_2, δ_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 size $\delta_1 > w_l > \delta_2$ 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 8 states that, among groups that inbreed and have a nonnegligible relative size, the expected Coleman index decreases as size grows. Second, Proposition 9 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 **8** *Consider any two groups, l and l', whose relative sizes are bounded away from 0 and 1 (i.e. there exists some $\vartheta > 0$ such that $1 - \vartheta \geq w_{l'} > w_l \geq \vartheta$). For* any *$\varpi > 0$, there exists some $\hat{n}$ such that if $n \geq \hat{n}$ and $w_{l'} - w_l \geq \varpi$, then $\mathbb{E}[\tilde{C}_l] > \mathbb{E}[\tilde{C}_{l'}]$.*

PROPOSITION **9** *There exist some $\hat{n}$ and $\delta_1 > \delta_2 > 0$ such that if $n \geq \hat{n}$, then any group l with relative size $1 - \delta_2 \geq w_l \geq 1 - \delta_1$ has $\mathbb{E}[\tilde{C}_l] < 0$.*

5 Further discussion

In this section, we complement our preceding analysis with a discussion of some other interesting features of the theory and the evidence. First, in Subsection 5.1, we focus on cross-group ties. At a theoretical level we consider the contrasting implications of one and two-sided setups, while at the empirical level we compare them in our high-school and marriage contexts. Second, in Subsection 5.2, we study the effect of total population size on homophily, focusing in particular on its effect on the Coleman index. Finally, in Subsection 5.3 we consider how group size affects the total number of successful meetings. In the latter two subsections, our empirical discussion restricts attention to the high-school data on friendship.

5.1 Cross-group ties

In this subsection we look at ties across different groups, and attempt to understand how their empirical distribution may be driven by some of the features of the framework – in particular, by whether the meeting framework is one- or two-sided.

As explained at the outset of Subsection 4.1, we can decompose the total number of ties of every group l by the group l' of their destination, giving rise to the corresponding magnitudes $(m_{ll'})_{l' \neq l}$. So, for instance, in the case of friendship ties, we may record the number of Black friends, or Hispanic friends, or Asian friends among the total number of friends enjoyed by White individuals. Then, if compute the corresponding proportions (over the total number of friends) for each other group with their respective frequencies in the population we arrive at what we will call the *excess representation* across various racial groups, defined by

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

where recall from Subsection 4.1 that $m_l \equiv \sum_{l'=1}^{q} m_{ll'}$ denotes the number of *total* matches of agents of type l.

The predictions of our model on cross-group ties crucially depend on which setup one adopts. In the one-sided variant, outbreeders meet agents in the whole pool, and, if meeting is uniform, find each group at rates that follow the relative sizes of these groups. In contrast, the two-sided variant 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. Thus, since outbreeding groups are relatively small (they are those whose size falls below the equilibrium threshold), the model predicts that cross-group matches are primarily formed among agents of small groups.

To find evidence of these predictions, we look at the relative representation of small groups in the friendships and marriages of other small groups. Figure 4 records the case of specific thresholds, set at 80 people for friendships and at 100 people for marriages. These cases are shown for illustrative purpose, and qualitatively similar pictures obtain when we fix different small thresholds. In the figure, for each group of relative size $w_l \in [0, 1]$, we associate $\sum_{\{l':n_{l'} \leq x\}} \Delta_{ll'}$, i.e. the aggregate excess representation of all those groups l' whose size amounts to no more than x people, where $x = 80$ for friendships and $x = 100$ for marriages.

The right panel of the picture suggests that a positive excess representation of "small" groups is a feature of the marriages of very small groups only, while 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. In our two-sided version of the model, the critical size suggested by these pictures is marked by

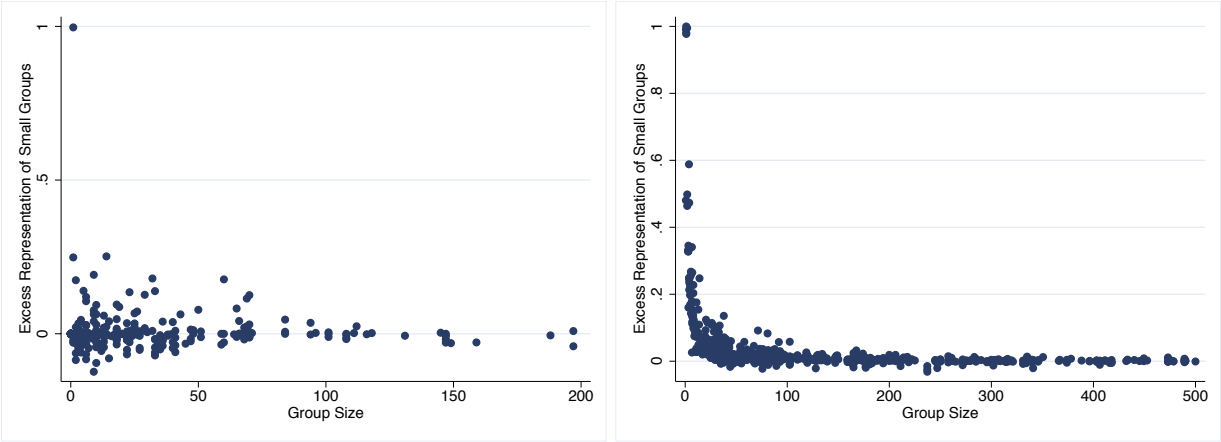

Figure 4: Inter-ethnic Ties with Small groups: Friendships (left) and Marriages (right).

the threshold beyond which there is a switch from an outbreeding to an inbreeding strategy. We do not find clear evidence of a similar trend for friendships. This provides some heuristic basis to conjecture that while a two-sided formulation may be more appropriate to model marriages, a one-sided approach may be adequate for friendships.

5.2 School size and homophily

School size (in terms of total number of students) has been shown by Currarini, Jackson and Pin (2010) to significantly affect the homophilous bias in the friendships of students of all ethnic groups. In particular, larger schools (with more than 1000 students) are shown to display uniformly larger Coleman indices across all groups' sizes than smaller schools (less than 1000 students). This increase in homophily is illustrated in Figure 5, and has been shown in that paper to be statistically significant. Also, this increase has been there shown to mostly reflect a difference in the meeting opportunities that students face in small versus large schools, which is interpreted as evidence of improved opportunities for self-segregation in larger schools, in the form of national societies and other race-segregated activities. These arguments are close in spirit to the main driving force of the present paper, evoking a minimal size of groups for certain "inbreeding" activities to be at work. In fact, as we argue here-below, the present setting provides a formal argument in support of these intuitions.[21]

Let τ 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 number of agents in the group, and is independent of the size n of the network for large n. In particular, this threshold is not defined in terms of the relative size of groups (that is, their fraction of the total population), which is measured

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

on the x-axis of Figure 5. As the number of students in the school (the parameter n in our model) increases, to any given relative group size w there corresponds a larger absolute size (that is, a larger number of group members). Denoting by $w(\tau, n)$ the relative group size that corresponds to the τ threshold for total population n, this implies that $w(\tau, n)$ is decreasing in n. So, as population increases from n to n', those groups with relative size 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 pattern can explain the shift in the relation between Coleman index and relative group size that we observe in Figure 5. This shift is substantial for small and medium sized groups, and vanishes for very large groups. Indeed, in the sample of smaller schools, observations with very small relative size are most likely to refer to groups with size below the threshold τ. Therefore, as we shift attention to the sample of larger schools, we find a higher extent of inbreeding behavior. For observations corresponding to medium relative size, the increase in inbreeding is less significant since many of the observations among small schools must correspond to groups that are already above the threshold. Finally, for observations with large relative size, no significant change is observed because most groups should be inbreeding, both in the sample of small and large schools.

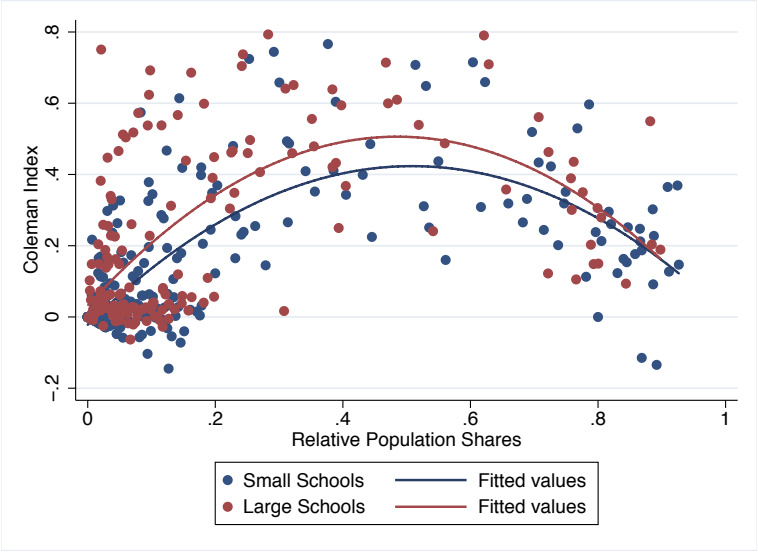

Figure 5: Coleman Index in Friendships: Small Schools (< 1000) vs. Large Schools (> 1000).

5.3 The effect of group size on popularity in high schools

Another important empirical regularity uncovered by CJP is that groups covering larger fractions of the school's population make more friends on average (see Figure 6 below). Under the assumption that the rate of encounters is unaffected by the size of the pool, the authors trace this pattern of

total friendships to a bias in preferences in favor of same type friends. The main channel through which preferences affect total matches in their model is the choice of meeting intensity, so that larger groups, facing better prospects in terms of type-mix of friends, devote more time to it. This aspect is absent from our model, where meeting intensity is exogenous and constant across groups.

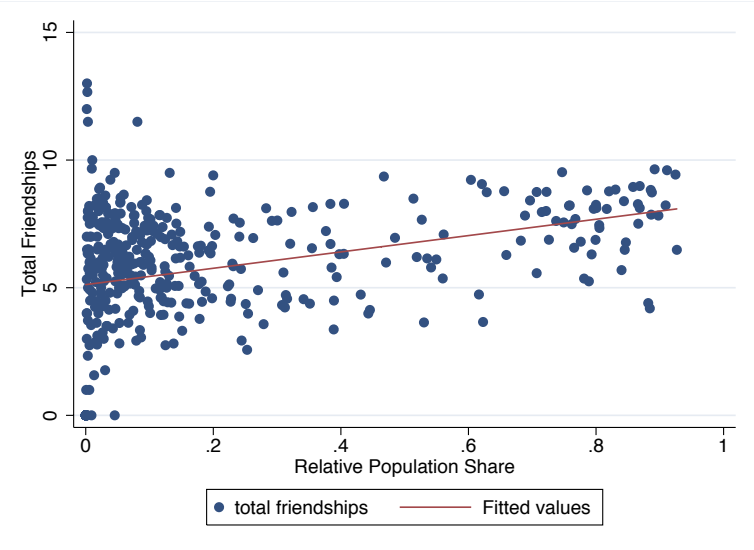

Figure 6: Total Friendships by Relative Group Size.

In the context of our model, group size can still affect the total number of matches if some (small) groups experience redundancies in their meeting possibilities at equilibrium. This is always the case in the presence of the meeting friction r_I, which "forces" members of all groups to direct some effort towards their own group only. Being exerted in pools of different sizes, this effort yields a larger expected number of friends to members of larger groups.

To identify this source of variability in groups' total friends, we regress the total number of friends on group size, controlling for the effect of relative population shares, which presumably contains, as argued above, elements of type-biased preferences. We perform this exercise on different ranges of group sizes, to capture the fact that the effect of group size should be larger for small groups, and presumably disappear for larger groups. We obtain the following results (* = 95% confidence; ** = 99% confidence):

0	w_l	n_l	constant
$n_l < 50$	3.20	0.07**	3.85**
$n_l < 100$	2.43*	0.03**	4.40**
$n_l > 100$	2.98**	0.00	5.27**
$n_l > 300$	3.53**	0.00	5.39**

While the ranges of group size used above are arbitrary, similar results obtain for minimal size threshold different from 100. The consistent trend is that the $p - value$ relative to n_l decreases with n_l, and significance is lost at around $n_l = 200$. The opposite holds for w_l, whose $p - value$ decreases with n_l, and the coefficient of w_l becomes significant at 99% level at around $n_l = 150$. This evidence is consistent with the effect of group size theorized in our model. Enlarging group size has a significant effect for very small groups, for which the enlargement yields a relevant relaxation of the meeting constraints imposed by a small size. This effect vanishes for larger groups, for which the effect of relative population shares takes over.

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. The 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 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 regularities observed on the pattern of in-group and cross-group ties, as well as with the nonmonotonicity displayed by the Coleman homophily index.

Homophily, however, is a complex and multifaceted phenomenon, so the purpose of our paper was not to construct a comprehensive model of it. Rather, our primary objective has been to propose a microfoundation of the meeting bias that, despite is simplicity, is consistent with many of the observed empirical patterns. We believe that our model highlights a very basic force underlying homophily and future analysis of the phenomenon should explicitly take in into account. Among the many set of issues that this future research should address, we would like to single out the need to allow for 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 interaction – 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 interaction/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] Bala, V. and S. Goyal (2000), "A Noncooperative Model of Network Formation," *Econometrica* 68, no. 5, 11811229.

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

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

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

[6] Currarini, S., M.O. Jackson, and P. Pin (2009) "An Economic Model of Friendship: Homophily, Minorities and Segregation," *Econometrica* 77, No. 4, 1003–1045.

[7] 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.

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

[9] Franz, S., M. Marsili and P. Pin (2008), "Observed choices and underlying opportunities", mimeo.

[10] Giles, M. W. (1978), "White Enrollment Stability and School Desegregation: A Two- Level Analysis" *American Sociological Review* 43, 2448-64.

[11] Golub, B. and M. O. Jackson (2011): "How homophily affects the speed of learning and best-response dynamics," mimeo.

[12] Lazarsfeld, P.F. and R.K. Merton (1954): "Friendship as a social process: a substanive and methodological analysis," in M. Berger (ed.), *Freedom and Control in Modern Society*, New York: Van Nostrand.

[13] Marsden, P.V. (1987): "Core Discussion Networks of Americans," *American Sociological Review* **52**, 122-313.

[14] Marsden, P.V. (1988): "Homogeneity in Confiding Relations," *Social Networks* **10**, 57-76.

[15] McPherson, M., L. Smith-Lovin and J. M. Cook (2001): "Birds of a Feather: Homophily in Social Networks", *Annual Review Sociology* 27, 415-44.

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

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

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

Appendix

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 γ 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:

$$\pi_O(n_l) \equiv V(\eta) - c - \delta(n), \tag{9}$$

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

$$\pi_I(n_l) \equiv \sum_{\nu_i=0}^{n} P_{\eta,n_l}(\nu_i)\ V(\nu_i). \tag{10}$$

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 imples 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 ν distinct elements from η draws with replacement out of a set of size θ. It is enough to show that, for all η and θ, 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 $\overrightarrow{P}_{\theta+1}(\nu;\eta)$:

$$\frac{\overrightarrow{P}_{\theta}(\nu;\eta)}{\overrightarrow{P}_{\theta+1}(\nu;\eta)} = \frac{\binom{\theta}{\nu} \sum_{j=0}^{\nu} (-1)^j \binom{\nu}{j} \left(\frac{\nu-j}{\theta}\right)^{\eta}}{\binom{\theta+1}{\nu} \sum_{j=0}^{\nu} (-1)^j \binom{\nu}{j} \left(\frac{\nu-j}{\theta+1}\right)^{\eta}} \tag{11}$$

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 θ and ν, the ratio $\frac{\overrightarrow{P}_{\theta}(\nu;\eta)}{\overrightarrow{P}_{\theta+1}(\nu;\eta)}$ is decreasing in ν. Since these are probability distributions, we conclude that these exists $\bar{\nu}$ such that $\frac{\overrightarrow{P}_{\theta}(\nu;\eta)}{\overrightarrow{P}_{\theta+1}(\nu;\eta)}$ 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,

$$\pi_I(n_l + 1) - \pi_I(n_l) > 0. \tag{12}$$

Let now τ^* be the lowest integer such that

$$\pi_I(\tau) \geq V(\eta) - c. \tag{13}$$

Then, both if (13) 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 τ^* 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 Θ 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 η independent draws with replacement out of the set Θ. Denoting by θ and l, the cardinalities of the sets Θ and L respectively, that expected number is equal to

$$(\theta - l) \cdot q_\theta(\eta) \tag{14}$$

where

$$q_\theta(\eta) = \left(1 - \left(\frac{\theta - 1}{\theta}\right)^\eta\right). \tag{15}$$

is the probability that an agent in the set Θ is found by means of η 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 θ when meeting is two-sided and both this agent and all the others obtain η draws. Its expected value can be computed by adding the expected number of agents who are met "actively" 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 met "passively" through the draws of others, 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:

$$\mathbb{E}\left[\tilde{\nu}_\leftrightarrow(\theta)\right] = \sum_{l=0}^{\theta} \binom{\theta}{l} q_\theta(\eta)^l (1 - q_\theta(\eta))^{\theta - l} \cdot l + \sum_{l=0}^{\theta} \binom{\theta}{l} q_\theta(\eta)^l (1 - q_\theta(\eta))^{\theta - l} \cdot (\theta - l) \cdot q_\theta(\eta). \tag{16}$$

Now note that by factoring the term $\theta \cdot q_\theta(\eta)$ in the second summatory in (16) we can write this sum as follows:

$$\theta \cdot q_\theta(\eta) \sum_{l=0}^{\theta} \binom{\theta}{l} q_\theta(\eta)^l (1 - q_\theta(\eta))^{\theta - l} - q_\theta(\eta) \sum_{l=0}^{\theta} \binom{\theta}{l} q_\theta(\eta)^l (1 - p(\eta, \theta))^{\theta - l} l. \tag{17}$$

and that the second term of (17) 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 (16) 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 θ 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 τ^* 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

$$\pi_O(n_l) \equiv V(\eta) - c - \delta(n), \tag{18}$$

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 τ^* as the only equilibrium threshold in this case, which completes the proof of the result. ■

Proof of Proposition 1

Consider any two types, l and l', with given finite group sizes $n_{l'} > n_l$. As in the text, denote by $m(s,\theta) \equiv \mathbb{E}[\tilde{\nu}(x,\theta)]$ the expected number of distinct meetings obtained when the number of draws is x and the pool size is θ. Suppose first that both groups are inbreeding. Then, for large n, the expected homophily indices for each group can be approximated arbitrarily well as follows:

$$\mathbb{E}\left[\tilde{H}_k\right] \simeq \frac{m(\eta + r_I, n_k)}{m(\eta + r_I, n_k) + m(r_O, \infty)} \qquad (k = l, l'),$$

which yields the desired conclusion since, by Lemma 2, $m(\eta + r_I, n_{l'}) > m(\eta + r_I, n_l)$. If both groups are outbreeding, then the corresponding expression is:

$$\mathbb{E}\left[\tilde{H}_k\right] \simeq \frac{m(r_I, n_k)}{m(r_I, n_k) + m(\eta + r_O, \infty)} \qquad (k = l, l'),$$

which leads to the same conclusion as before since $m(r_I, n_{l'}) > m(r_I, n_l)$, again by Lemma 2.

Finally, consider the case where groups l and l' do not make the same breeding decision. Then, due to the threshold property of the equilibrium, it must be that group l outbreeds and group l' inbreeds. Thus Lemma 2 still yields the desired conclusion as

$$\begin{aligned} &m(\eta + r_I, n_{l'})\, m(\eta + r_O, \infty) - m(\eta + r_I, n_l)\, m(r_O, \infty) = \\ &m(\eta + r_I, n_{l'})\, (\eta + r_O) - m(\eta + r_I, n_l)\, r_O > 0 \end{aligned}$$

,

which implies that, for large enough n, $\mathbb{E}\left[\tilde{H}_{l'}\right] - \mathbb{E}\left[\tilde{H}_l\right] > 0$. ■

Proof of Proposition 2

It is enough to note that, for any inbreeding group l whose size n_l is bounded above, the $r_I + \eta$ inbreeding draws can yield distinct meetings $q \in \{1, 2, \ldots, r_I + \eta\}$, all of these possibilities with a probability that is bounded above zero, independently of n. ■

Proof of Proposition 3

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. We can then write:

$$\begin{aligned} \Delta &= \left[\frac{m(r_I, n_l')}{m(r_I, n_l') + m(r_O + \eta, \infty)} - \frac{n_{l'}}{n}\right] - \left[\frac{m(r_I, n_l)}{m(x, n_l) + m(r_O + \eta, \infty)} - \frac{n_l}{n}\right] \\ &= \frac{m(r_I, n_l')}{m(r_I, n_l') + m(r_O + \eta, \infty)} - \frac{m(r_I, n_l)}{m(r_I, n_l) + m(r_O + \eta, \infty)} - \frac{1}{n}. \end{aligned} \tag{19}$$

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, Δ is strictly positive. ■

Proof of Proposition 4

Consider any type l and, for large n, approximate its expected homophily index as follows:

$$\mathbb{E}\left[\tilde{H}_k\right] \simeq \frac{m(\eta + r_I, n_l) + m(r_O, \infty) \cdot w_l}{m(\eta + r_I, n_l) + m(r_O, \infty)}.$$

If the relative size of the respective group, w_l, remains bounded above zero, by relying on usual considerations, we may write:

$$\mathbb{E}\left[\tilde{H}_k\right] = \frac{\eta + r_I + r_O \cdot w_l}{\eta + r_I + r_O} + \delta(n)$$

with $\delta(n) \to 0$ as $n \to \infty$. Thus the expected excess homophily of type l can be written as

$$\mathbb{E}\left[\tilde{H}_k\right] - w_l = \frac{\eta + r_I + r_O}{\eta + r_I + r_O} + \left(1 - \frac{r_O}{\eta + r_I + r_O}\right) w_l + \delta(n)$$

so that, combining it with the corresponding magnitude for any other group l' with size $w_{l'} > w_l$ we conclude that

$$\left\{\mathbb{E}\left[\tilde{H}_{l'}\right] - w_{l'}\right\} - \left\{\mathbb{E}\left[\tilde{H}_l\right] - w_l\right\} = \left(1 - \frac{r_O}{\eta + r_I + r_O}\right)(w_{l'} - w_l) + \delta'(n) - \delta(n)$$

which, making $\alpha = (1 - \frac{r_O}{\eta + r_I + r_O})$, yields the desired conclusion. ■

Proof of Proposition 5 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 ϵ, 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 ϵ is chosen small enough. ■

Proof of Proposition 6 Consider any two outbreeding groups l and l' with $n_l < n_{l'}$. Since, obviously, $1 - w_{l'} < 1 - w_l$, in order to prove that $\mathbb{E}[C_l] < \mathbb{E}[C_{l'}]$ it is enough to show that the expected excess homophily is ordered in the same direction for the two groups, i.e.

$$\mathbb{E}\left[H_l\right] - w_l < \mathbb{E}\left[H_{l'}\right] - w_{l'}.$$

But this is precisely what was established by Proposition 3, so the proof is complete. ■

Proof of Proposition 7 A preliminary observation is that, if n is large enough, then since $w_l = \frac{n_l}{n}$ is bounded away from zero by δ_2 it must be that $n_l \geq \tau^*$ (where τ^* 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 δ_3 ensures that the term $\frac{\eta+r_I}{\eta+r_I+r_O}$ is arbitrarily close to 1. Thus, by choosing δ_1 small enough, the expected homophily $\mathbb{E}[\tilde{C}_l]$ can be made arbitrarily close to 1, as desired. ■

Proof of Proposition 8 Consider two groups, l and l', whose relative sizes 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 τ^* 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:

$$\frac{\frac{m(\eta+r_I,\infty)}{m(\eta+r_I,\infty)+m(r_O,\infty)} - w_l}{1-w_l} > \frac{\frac{m(\eta+r_I,\infty)}{m(\eta+r_I,\infty)+m(r_O,\infty)} - w_{l'}}{1-w_{l'}} \tag{20}$$

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 (20) holds if, and only if, the difference $w_{l'} - w_l$ is bounded above zero, as claimed. ■

Proof of Proposition 9 Given η, 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 δ_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. ■