



## Search and Homophily in Social Networks

**Sergio Currarini**

*Università Ca' Foscari di Venezia*

**Fernando Vega Redondo**

*European University Institute*

First Draft: 23.10.2010

### Abstract

We study the formation of social ties among heterogeneous agents in a model where meetings are governed by agents' directed search. The aim is to shed light on the important issue of homophily (the tendency of agents to connect with others of the same type). The essential contribution of the model is to provide a basic microfoundation for the opportunity/meeting biases that, as the literature highlights, are a crucial element of the phenomenon. Under the assumption that search is more effective in large pools, the equilibrium is characterized by a threshold in terms of group size: large groups only search among similar agents while smaller groups search in the whole population. This threshold behavior is consistent with the empirical evidence observed in a range of social environments such as high school friendships and interethnic marriages. And assuming that search is subject to small frictions, it also generates the bell-shaped form of the so-called Coleman index observed in the data. Other implications of the model supported by the evidence concern the pattern of cross-group ties among small groups, the linearity of excess homophily for large groups, and the positive effect on it of overall population size.

### Keywords

Homophily, search, social networks, segregation.

### JEL Codes

D7, D71, D85, Z13

### Address for correspondence:

**Sergio Currarini**  
Department of Economics  
Ca' Foscari University of Venice  
Cannaregio 873, Fondamenta S.Giobbe  
30121 Venezia - Italy  
Phone: (+39) 041 2349133  
Fax: (+39) 041 2349176  
e-mail: s.currarini@unive.it

*This Working Paper is published under the auspices of the Department of Economics of the Ca' Foscari University of Venice. Opinions expressed herein are those of the authors and not those of the Department. The Working Paper series is designed to divulge preliminary or incomplete work, circulated to favour discussion and comments. Citation of this paper should consider its provisional character.*

The Working Paper Series  
is available only on line  
([www.dse.unive.it/publicazioni](http://www.dse.unive.it/publicazioni))  
For editorial correspondence, please contact:  
[wp.dse@unive.it](mailto:wp.dse@unive.it)

Department of Economics  
Ca' Foscari University of Venice  
Cannaregio 873, Fondamenta San Giobbe  
30121 Venice Italy  
Fax: +39 041 2349210



# Search and Homophily in Social Networks

Sergio Currarini

Fernando Vega Redondo\*

October 23, 2010

## Abstract

We study the formation of social ties among heterogeneous agents in a model where meetings are governed by agents' directed search. The aim is to shed light on the important issue of homophily (the tendency of agents to connect with others of the same type). The essential contribution of the model is to provide a basic microfoundation for the opportunity/meeting biases that, as the literature highlights, are a crucial element of the phenomenon. Under the assumption that search is more effective in large pools, the equilibrium is characterized by a threshold in terms of group size: large groups only search among similar agents while smaller groups search in the whole population. This threshold behavior is consistent with the empirical evidence observed in a range of social environments such as high school friendships and interethnic marriages. And assuming that search is subject to small frictions, it also generates the bell-shaped form of the so-called Coleman index observed in the data. Other implications of the model supported by the evidence concern the pattern of cross-group ties among small groups, the linearity of excess homophily for large groups, and the positive effect on it of overall population size.

*Keywords:* Homophily, search, social networks, segregation.

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

---

\*Currarini: Dipartimento di Scienze Economiche, Università di Venezia and School for Advanced Studies in Venice (SSAV). Email: s.currarini@unive.it; Vega Redondo: European University Institute (Florence) and Instituto Valenciano de Investigaciones Económicas. Email: Fernando.Vega@eui.eu. The authors wish to thank Matt Jackson, Paolo Pin, Marcel Fafchamps and the participants to seminars at the University of Maastricht, European University Institute, University of Bristol, Econometric Society 2010 World Meeting.

# 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 and other aspects of social communication flow on the network of social contacts, and to what extent distance in characteristics translates into distance in the network. It is therefore important to understand more about the generative process of homophilous social networks, and how the attitudes of agents and their meeting opportunities concur in determining the observed mix of social ties.

One first determinant of this mix is the distribution of agents across homogeneous groups. As pointed out by Blau (1977), the relative sizes of groups affect the distribution of in- and cross-group links by affecting the meeting opportunities of agents.<sup>1</sup> If ties were formed uniformly at random, 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, which affects the distribution of in-group ties even in the absence of biases in agents' attitudes towards dissimilar agents or in their meeting opportunities.

What is most striking 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, interethnic integration, it is crucial to have a proper understanding of such endogenous forces at work.<sup>2</sup>

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.<sup>3</sup> Among these, the closest to our concerns is

---

<sup>1</sup>The notion that certain opportunities of encounters enhance the probability of tie formation is already present in Alport's (1954) contact theory

<sup>2</sup>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, will be better served by dispersion rather than concentration.

<sup>3</sup>See, among others, the papers by Currarini, Jackson and Pin (2009, 2010a, 2010b) (discussed below in some detail), and Golub and Jackson (2010) on the effect of homophily on information transmission, Buhai and van der

Currarini, Jackson and Pin (2009) – henceforth referred to as CJP – which proposes a model of tie formation that is tested against data on friendship networks obtained at racially heterogeneous American high schools. We shall often find it useful to refer to that paper as a way of motivating or understanding our present endeavour. One of its primary concerns is to disentangle the incidence of preference and meeting biases in the formation of social networks. And, in this respect, the key insight is that the patterns of homophily observed in friendship networks *cannot* be explained by an homophilous bias in *preferences* alone.<sup>4</sup> That is, they find that there must be some *meeting* bias in the sense that the probability of meeting own types is higher than their corresponding frequency in the search pool.

This 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. The key ingredients are as follows. Agents are partitioned in type-homogeneous groups, and derive positive utility from the number of ties they enjoy. Their single decision is to select the pool where to search for ties. Two options are available to them: either they restrict search to individuals of their own type – a decision we call *inbreeding* – or they extend search to the whole population – we label it *outbreeding*. This dichotomous dilemma must confront two sorts of conflicting incentives. On the one hand, outbreeding entails a cost, which we assumed fixed and may come from cultural, geographical, or linguistic barriers to accessing other types. But, on the other hand, inbreeding has the drawback that it limits the scope of search – specifically, we posit that smaller pools lead to a (stochastically) inferior set of search draws.<sup>5</sup>

The two previous considerations generate an interesting tension between inbreeding and outbreeding, which turns out to be resolved in a drastically different manner in large and small groups. 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 model is that the meeting bias arising at equilibrium does *not* hold across the whole population and just applies to large groups. Only the members of these groups, that is, find it optimal not to pay the cost of outbreeding, which leads them to meet agents of their own type alone.

Our model is abstract and axiomatic, which has the advantage that it applies to a wide range of different specific contexts. In the end, however, the outcome at equilibrium hinges upon how agents’ breeding choices shape their respective search pool. We study two polar scenarios: one-sided and

---

Leij (2009) and Patacchini and Zenou (2008) on homophily in the labour market, and Barr, Dekker and Fafchamps (2009) for a field experimental of gender homophily.

<sup>4</sup>In the framework studied by CJP, the preference bias plays the important role of affecting the *number* of friends per capita.

<sup>5</sup>We shall provide below concrete illustrations in which such an assumption is well justified. For example, this happens when small pools suffer from various forms of redundancy in search, or lack variety due to strong correlations in tastes.

two-sided. In the one-sided variant of the model, outbreeding agents search in the population at large (outbreeding or not); this is representative of situations in which agents can unilaterally decide to activate a connection, as in some instances of information acquisition (e.g. internet browsing). By contrast, in the two-sided variant, 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*.

To test empirically our model, we take it to data gathered in two quite diverse contexts: high school friendships from the widely used AddHealth dataset, and interethnic marriages from U.S. census data.<sup>6</sup> First we show that our model matches the behavior of the Coleman Index (a normalized measure of homophily in excess of population shares) that CJP report for U.S. high school friendships. This index depends nonlinearly on group size, with maximal values for middle-sized groups and low positive values for small and large groups. Here we show that this index shows a similar pattern for U.S. marriages.<sup>7</sup> In this respect, our model can be regarded as providing some simple but plausible microfoundation to the meeting bias that is directly postulated by CJP to match the empirical evidence.

Our model also yields a good number of further empirical predictions that are intimately linked to the specific breeding behavior of agents that is induced by its equilibrium. We start by focusing on three of them.

1. A first one concerns the share of in-group ties as a function of groups’ size. The prediction is that this share should display an abrupt jump up to some small threshold level for group size, beyond which it should stay constant at one (since those groups are uniformly inbreeding).
2. Another prediction concerns the difference between the share of in-group ties and the population share of groups. This difference can be thought of as the homophily in excess of what theoretically expected *via* uniform random assortment. The model predicts that this difference should be linearly decreasing in the population share for inbreeding groups.
3. A third prediction concerns the pattern of cross-group ties. When link-formation is two-sided, our model predicts that a cross-group tie should occur if and only if both agents are outbreeders. Thus, in equilibrium, cross-group ties must involve members of small groups alone.

We find that the first two predictions are matched well by both our friendships and marriages datasets. Empirical support is also found for the third prediction, the evidence being stronger for inter-ethnic marriages than for friendships.<sup>8</sup>

---

<sup>6</sup>See section 4 for a detailed description of these datasets.

<sup>7</sup>See also Rogers and Bramoullé (2009).

<sup>8</sup>The larger variance observed in the case of friendships may be due to a stronger role for preferences in this case.

Finally, it is worth highlighting that our model also sheds light on, and complements, other empirical regularities identified by the previous literature. As an illustration, we list the following two:

1. The model predicts that, conditional on the relative shares of each group, the larger is the overall population the (weakly) higher is number of inbreeding groups.<sup>9</sup> When observations are pooled separately for large and small populations, one then expects higher Coleman indices in large populations than in small ones. This effect of population size on homophily was documented by Currarini, Jackson and Pin (2010a) for high school friendships.
2. The existence of a significant positive relation between the population shares of groups and total ties was documented by CJP, and explained as a consequence of the higher search intensities exerted by the members of groups that cover a larger fraction of the overall population. In our model, where search intensity is fixed, a similar implication obtains for small groups (a manifestation of the lower effectiveness of search conducted in such groups). But for large groups, our model predicts no significant effect of size, which is a feature that we find corroborated by the data.

To end this introduction, we outline the ensuing contents of the paper. Section 2 describes the search model, with the main axioms, illustrations, and the search game. Section 3 characterizes the equilibria of the game. Section 4 studies the implication of the model for homophily. It first describes the empirical evidence derived from friendships networks in U.S. high schools and U.S. marriages. It then discuss the implication of the search equilibrium for homophily, showing that the resulting characterization fits well the empirical evidence. Section 5 concludes the paper.

## 2 The Search 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  $l = 1, 2, \dots, q$ . Each agent  $i$  devotes  $\eta$  units of time or effort to search for suitable matches among the agents in  $N$ . The sole decision every agent takes is how to allocate time between search among agents of her own group and search within the whole population (including her group). We will refer to the first type of search 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., travelling, learning a language, or changing one's habits).

---

<sup>9</sup>In our model, the reason for this conclusion is obvious. The threshold above which groups inbreed is defined in terms of the *absolute* size of groups. The argument follows from the observation that for any given relative share, a rise in population size can never make the size of a former inbreeding group fall below this threshold.

The inbreeding/outbreeding decisions taken by all agents define the pools in which each of them searches, and these pools in turn determine payoffs by affecting the matching outcomes. In what follows, we address these considerations in the following order: firstly, we describe how the pool size affects the probability distribution over the number of matchings; secondly, we discuss how these matching probabilities determine payoffs and thus shape agents' preferences and decisions; thirdly, we explain how agents' in-/out-breeding decisions conform their respective search pool; and finally, we shall combine all former features to provide a formal specification of the search game that underlies our analysis.

## 2.1 Search pool and matching success

Let  $\Theta$  be the search pool faced by any given agent, i.e. the set of all those other individuals with whom she can be matched through search. For any level of search intensity  $\eta$ , define by  $\Pi(\eta, \Theta)$  the probability distribution over subsets of  $\Theta$  specifying, for every  $S \subset \Theta$ , the probability  $\Pi(\eta, \Theta)(S)$  that the set of distinct matches of the agent coincides with  $S$ . For any given set  $S$ , denote by  $|S|$  the cardinality of this set. We start by positing a technical condition of continuity.

**CON (Continuity)** Given any  $\varepsilon > 0$ , there exists some  $\delta > 0$  such that, for all  $S \subset \Theta$ ,

$$\frac{|S|}{|\Theta|} < \delta \Rightarrow [\forall T \subset \Theta, T \cap S \neq \emptyset \Rightarrow \Pi(\eta, \Theta)(T) < \varepsilon].$$

The previous condition is in the spirit of the common assumption of absolute continuity. It demands that subsets of agents of arbitrarily small relative size in a large pool must also have an arbitrarily small probability of being found through search. We postpone a discussion of the role and of the implications of this assumption for our results until section 4, where the equilibrium patterns of in-group and cross-groups matches are examined.

The essential features of the matching process are assumed independent of the identity of the agents and therefore simply associated to the search intensity  $\eta$  exerted by an agent and the size of her matching pool  $\theta \equiv |\Theta|$ . These two variables define what we call a *search profile*  $(\eta, \theta)$ , to which we associate a random variable  $\nu(\eta, \theta) \in \mathbb{N}$  that governs the number of potential matches associated with the profile  $(\eta, \theta)$ . We assume that  $\nu(\eta, \theta)$  is continuous in  $\eta$  (with the topology of weak convergence) and has  $\nu(0, \theta)$  concentrated in zero. In general, we shall posit that every such potential match is suitable (or successful) only with some probability  $p \in (0, 1]$ , reflecting the (stochastically independent) event that any two agents are indeed compatible.

We assume that  $\nu(\eta, \theta)$  is continuous in  $\eta$  (with the topology of weak convergence) and has  $\nu(0, \theta)$  concentrated in zero. But the key assumption that will underlie our analysis is a condition of monotonicity on how the number of potential matches depend on the size  $\theta$  of the search pool. This general idea will be contemplated in two different forms, one stronger than the other.

**SM (Strong Monotonicity)** Let  $(\eta, \theta) \geq (\eta', \theta')$ . Then, the induced variable  $\nu(\eta, \theta)$  strictly dominates  $\nu(\eta', \theta')$  in the first-order stochastic sense. Moreover, the variable  $\nu(\eta, \infty) \equiv \lim_{\theta \rightarrow \infty} \nu(\eta, \theta)$  has a well defined distribution for all finite positive values of  $\eta$ .

**WM (Weak Monotonicity)** Let  $(\eta, \theta) \geq (\eta', \theta')$ . Then, the induced expected number of potential matches satisfies  $\mathbb{E}[\nu(\eta, \theta)] > \mathbb{E}[\nu(\eta', \theta')]$ . Moreover, the value  $\mathbb{E}[\nu(\eta, \infty)] \equiv \lim_{\theta \rightarrow \infty} \mathbb{E}[\nu(\eta, \theta)]$  is well defined for all finite positive values of  $\eta$ .

The above two axioms are intended to capture a common feature that arises in a number of different search setups, even if the “micro details” underlying the specific mechanism at work may well differ (see the illustrations in the next subsection). Both express the key idea that search is more effective in larger pools. In some cases, First-Order Stochastic Dominance (FOSD) is too demanding a criterion, and we shall resort to the weaker axiom formulated in terms of expectations rather than full distributions.

## 2.2 Illustrations of the monotonicity axioms

We present three illustrations of search environments where the postulated monotonicity axioms hold. In the first context, search is carried out through random draws with replacement, thus becoming more effective in larger pools because size reduces the likelihood of redundancies (i.e. the probability that a previously found individual is met again). The second setup assumes that new ties are formed through the existing social network, which again has smaller groups display a larger probability of redundancy due to the entailed higher clustering. Finally, the third scenario illustrates how size monotonicity can arise as a consequence of a preference for variety when characteristics are correlated with groups.

### 2.2.1 Search through random draws with replacement

Consider the following search process. Each agent in a population makes a given number  $\eta > 1$  of independent draws with replacement out of a population of size  $\theta < n$ . Within such a general set-up we contemplate two different variants.

In the first one, search is of a one-way flow nature, in the sense that the matches actually enjoyed by an agent can only come from her own draws (and not, for instance, from the draws of other agents who find her). This may represent information acquisition on the internet, where an agent only gets benefits from the sources of information she finds, or traditional marriage markets, in which one side choose and propose to the other side, which is passive in the search process (in this latter example it is natural to have  $\theta = n/2$ ).

In the second variant, search is of the two-way flow type, and draws have bidirectional implications on the two agents involved. Thus, in this case, an agent gets the benefit of new partners either by finding them (i.e. as the outcome of her own search draws) or by being found (i.e. as the result

of others' search). This is the natural formulation if, say, we think of ties as reflecting friendship or we want to model marriage formation in modern societies. In these cases, the direction of search induces no strong asymmetries on the enjoyment of the tie by the two agents involved.

Consider first the one-way flow variant. The matches that are relevant for an agent are the distinct individuals she meets as an outcome of one of her  $\eta$  draws with replacement. For, as an example, finding the same piece of information a second time does not add anything to what the agent already knew. Let us denote by  $\nu_a(\eta, \theta)$  the associated random variable that determines the number of distinct draws in a pool of size  $\theta$ . The following result establishes that this context satisfies our strong monotonicity axiom.

**PROPOSITION 1** *The random variable  $\nu_a(\eta, \theta)$  satisfies axiom SM. Moreover, when the pool size grows large ( $\theta \rightarrow \infty$ ) the distribution converges to the degenerate Dirac distribution  $\nu_a(\eta, \infty) = \eta$ .*

The proof of this and of all other propositions are found in the appendix. The result of Proposition 1 is intuitive: in a larger pool, the probability of finding the same person repeatedly is lower, and so the number of distinct draws increases (in a stochastic sense). In the limit, as the population gets arbitrarily large, all  $\eta$  search draws are distinct.

Let us now turn to the two-way flow variant. Consider any agent  $i \in N$ . She makes  $\eta$  random draws with replacement out of the set  $N \setminus \{i\}$  while, at the same time, every  $j \in N \setminus \{i\}$  makes  $\eta$  random draws with replacement out of a set  $N \setminus \{j\}$ , which naturally includes  $i$ . Define the random variable  $\nu_{ap}(\eta, \theta)$  as the number of elements of the set  $N \setminus \{i\}$  which are either found at least once by  $i$ , or that find  $i$  at least once, or both.<sup>10</sup> The variable  $\nu_{ap}(\eta, \theta)$  is the union of distinct active and passive draws for  $i$ . The next proposition establishes that  $\nu_{ap}(\eta, \theta)$  satisfies the Weak Monotonicity axiom.

**PROPOSITION 2** *The random variable  $\nu_{ap}(\eta, \theta)$  satisfies the WM axiom.*

To see why Strong Monotonicity is too strong an assumption for this "two-way flow" setup, consider agent  $i$  and the effect on  $i$ 's matches of an enlargement of the pool to which  $i$  belongs. This has, as in the one-way flow case, a strong positive effect on the distribution of distinct draws. But, in addition to this effect, now a larger pool also has other implications on the passive matches that are more ambiguous. For, as the pool grows, not only does it happen that the set of searching agents who can find  $i$  grows (which again increases the probability that  $i$  is found). It also renders it *less* likely that  $i$  is found by any given agent in the pool. This last effect has negative implications on the number of distinct (passive) draws that player  $i$  receives, and is responsible for the weaker criterion we must use to rank the random variables in the two-way flow variant.

<sup>10</sup>Note that here the variable  $\theta$  is here an index of group (rather than pool) size, since agent  $i$  searches among  $\theta - 1$  agents, and  $\theta - 1$  agents search agent  $i$  among  $\theta - 1$  agents. Also note that the process described does *not* correspond a marriage problem, where the set of agents are actually partitioned in two subset within which agents never search one another. Our analysis, however, extends to that case, with the understanding that a growing pool size means that both sides of the marriage market grow in size.

### 2.2.2 Finding friends through friends

Consider a search environment in which each agent of group  $l$  is endowed with a set of "friends" belonging to the same group  $l$ . These friends are exogenously given, and can be used to search for additional friends, by randomly drawing among their own friends. Formally, let us construct a random network within each group by defining a set of neighbors for each agent.<sup>11</sup> If average degree is assumed constant across groups independently of size, then it is easy to see that the clustering of the networks prevailing in each group is inversely related to groups' size. This implies that the probability of search redundancy (that is, of finding an agent who is already a friend) is smaller in larger groups. Or, reciprocally, the probability of meeting a new friend through existing friends is larger, as suggested.

In this search environment, the choice of *inbreeding* can be identified with the choice of using the network to find new friends, since this choice would always lead to meeting people from one's own group. In contrast, *outbreeding* would correspond to the decision of relying on some anonymous global mechanism that induces a uniform probability distribution over all agents in the population. And, in this context, one can associate the cost of outbreeding to the additional search effort required to use of an anonymous matching technology rather than one used based on existing friends. This seems to well describe certain features of immigrants' social networks, where original social ties are prominently with other immigrants. In those cases, some significant cost (e.g. learning the dominant language, or adapting one's way of life) must be incurred in order to expand the set of friends beyond the original communities.

### 2.2.3 Taste for variety

Here we want to think of groups as sets of agents that are defined relative to some metric on relevant socio-economic characteristics, such as geography, resource endowments, kinship, race, or religion. For example, as in Dixit's (2003) model of trade, it may be assumed that agents are located on a circle, with their distance defined as the length of the shortest arc joining them. Then, a group simply consists of an arc of adjacent agents.

Next assume that each agent is randomly assigned some characteristic (different from the group type) according to some probability distribution over a given set. And suppose that these probabilities depend on distance. Specifically, let us postulate that the conditional probability that two given agents display a different characteristic grows with distance, i.e. distance tends to breed diversity. This is important because there are strong complementarities in, say, either production or preferences. Thus a match can be successful only if the two agents involved display a different characteristic.

Now suppose that agents' search cannot target others according to distance but they can decide only whether to search within their own group or outside it. And, as usual, we posit that there

---

<sup>11</sup>See, for example, Vega-Redondo (2007) for an account of random network models in social contexts.

is a cost to be paid to search outside one’s group. Then, if we think of the random variable  $\nu$  as counting the number of potentially successful matches (i.e. those involving agent with different characteristics), the contemplated size-dependent (weak) monotonicity follows from our assumptions. Specifically, the WM axiom applies since a larger group includes more ”variety” in expected terms.

### 2.3 Preferences over matching outcomes

We now go back to our stylized model of search, and describe agents’ preferences upon the number of matches that they enjoy. Denote this number by  $y_i$  for each agent  $i$ . Given the stochastically independent probability of success  $p$  of each potential match, the number of successful matches  $y_i$  is related to the number of potential matches  $\nu(\eta, \theta)$  according to a binomial distribution.

More specifically, let  $U : \mathbb{N} \cup \{0\} \rightarrow \mathbb{R}_+$  denote the von Neuman-Morgenstern (vNM) utility representing agents’ preferences over successful matches. Then, we can correspondingly define a vNM utility on the number  $\nu_i$  of potential matches of agent  $i$ . This function  $V : \mathbb{N} \cup \{0\} \rightarrow \mathbb{R}_+$  is defined by

$$V(\nu_i) = \sum_{y_i=0}^{\nu_i} \binom{\nu_i}{y_i} p^{y_i} (1-p)^{\nu_i-y_i} U(y_i). \quad (1)$$

We assume that  $U(0) = 0$  and that

$$U(y_i + 1) \geq U(y_i) \quad \text{for all } y \in \mathbb{N} \quad (2)$$

$$U(1) > U(0). \quad (3)$$

Thus we only demand that no successful draw is worse than one, but there may well be saturation beyond that point. This general framework accommodates several applications, from friendships to 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.

From an ex ante viewpoint, however, the number of potential matches is uncertain. We thus need to integrate (1) with the distribution over distinct/potential matches induced by a search profile  $(\eta, \theta)$ , to obtain the expected payoff associated with any such profile as follows:

$$\mathbb{E}_{\nu(\eta, \theta)} V(\nu_i) = \sum_{\nu_i=0}^n P_{\eta, \theta}(\nu_i) V(\nu_i). \quad (4)$$

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

### 2.4 The search game

We are now in a position to define the search game. In principle, the strategy of any agent  $i$  should determine the intensities  $\eta_I$  and  $\eta_O$  (with  $\eta_I + \eta_O = \eta$ ) to be devoted to inbreeding and outbreeding,

respectively. For simplicity, however, we restrict attention to a context where this decision is strictly dichotomous, so the agent simply decides whether to direct search towards members of her own group only ( $\eta_I = \eta$ ) or to extend search to all groups in an unbiased manner ( $\eta_O = \eta$ ). We call the first alternative "inbreeding" ( $I$ ), and the second "outbreeding" ( $O$ ). In the appendix – cf. Lemma 1 – we show that such a binary choice setup can be assumed without loss of generality under a mild additional axiom that postulates an additivity requirement on the random variable  $\nu$  when the pool is very large

A strategy profile of the search game can therefore be identified with a vector  $(s_i)_{i \in N} \in \{I, O\}^n$ . For convenience, we shall focus throughout on strategy profiles  $s$  that are group-symmetric, i.e. where  $s_i = s_j$  whenever  $i$  and  $j$  belong to the same group. Thus the population behavior can be fully described by the  $q$ -tuple  $\gamma \equiv (\gamma_1, \gamma_2, \dots, \gamma_q)$  that specifies the common choice  $\gamma_l \in \{I, O\}$  for every agent in the groups  $l = 1, 2, \dots, q$ .

Given any such profile  $\gamma$ , the induced matching outcome enjoyed by the individuals of each group  $l$  is determined by the random variable  $\nu(\eta, \theta_l(\gamma))$ , where  $\theta_l(\gamma)$  is the search pool faced by agents of group  $l$  under the strategy profile  $\gamma$ . As advanced, we shall distinguish two different scenarios determining how the pool of each group is shaped by the strategy profile  $\gamma$  through the mapping  $[\theta_l(\cdot)]_{l=1}^q$ .

The simplest scenario is *one-sided*, in that the search conditions enjoyed by any given agent exclusively depend on her own breeding decision,  $I$  or  $O$ . In particular, an outbreeding agent searches among all agents in the system. Thus if we denote by  $\theta_l(\gamma)$  the size of the search pool accessed in this context by group  $l$  when the (group-) strategy profile is  $(\gamma_1, \gamma_2, \dots, \gamma_q)$ , we have:

$$\theta_l(\gamma) = \begin{cases} n_l & \text{if } \gamma_l = I \\ n & \text{if } \gamma_l = O \end{cases} \quad (5)$$

where recall that  $n_l$  stands for the cardinality of group  $l$ . This formulation can be suitable to model situations in which outbreeding occurs by, say, taking the initiative in visiting the locations where other groups live or learning the languages they speak.

In contrast, a two-sided context is one where the search pool of any outbreeding group consists of those groups that have themselves chosen to outbreed. This gives rise to an alternative function  $\tilde{\theta}_l(\gamma)$  specifying the corresponding size of the search pool:

$$\tilde{\theta}_l(\gamma) = \begin{cases} n_l & \text{if } \gamma_l = I \\ \sum_{\{l': \gamma_{l'} = O\}} n_{l'} & \text{if } \gamma_l = O \end{cases} \quad (6)$$

This alternative scenario can be used to model situations in which, say, all outbreeders move to some fixed location ("downtown") where only outbreeders meet, or they learn a common language distinct from the one each of them originally speaks.

**REMARK 1** *It is worth emphasizing that the contrast between one- and two-sided matching considered here is conceptually different from the distinction between one-way and two-way flow contexts introduced in Subsection 2.2. The type of flow pertains to how payoffs originate from successful draws. Instead, the sidedness of ties concerns how matching is conducted (i.e. pools are formed) as a result of agents' breeding decision. Admittedly, despite the conceptual independence of these alternatives, some combinations appear to be most natural – e.g. two (resp. one) directional flows and two (resp. one) sided matchings. But other alternatives might be suitable descriptions as well of some situations. For example, two-sided matchings and one-way flows may reflect a context where outbreeding is achieved by visiting a common location (thus it is two-sided) and, once there, agents gather information from those whom they choose to contact and ask (thus payoffs are of the one-way flow type). The important point to bear in mind is that while the assumption on flows affects the distribution of successful matches in a given pool (and bears, for example, on the appropriate notion of monotonicity), their “sidedness” affects the way in which matching pools are defined.*

To sum up, the search game is fully defined as follows. The players are the population  $N$ , partitioned into  $q$  groups or types. For each individual  $i \in N$ , her strategy set  $S_i = \{I, O\}$  includes the inbreeding and outbreeding decisions. Under type-symmetry, any strategy profile  $\gamma \in \{I, O\}^q$  induces the size  $\theta_l(\gamma)$  of the matching pool faced by the individuals of each group  $l = 1, 2, \dots, q$ . This then defines the payoff function  $\pi_l(\gamma)$  of the representative agent of each group as

$$\mathbb{E}_{\nu(\eta, \theta_l(\gamma))} V(\nu_i) = \sum_{\nu_i=0}^n P_{\eta, \theta_l(\gamma)}(\nu_i) \sum_{y_i=0}^{\nu_i} \binom{\nu_i}{y_i} p^{y_i} (1-p)^{\nu_i-y_i} U(y_i),$$

where  $P_{\eta, \theta_l(\gamma)}(\cdot)$  denote the probabilities over the potential matches induced by the random variable  $\nu(\eta, \theta_l(\gamma))$ ,  $p$  is the probability that each potential match be successful, and  $U(\cdot)$  defines the utility of agents over successful matches.

### 3 Search Equilibrium

We now characterize the group-symmetric Nash equilibria of our search game. The central result is that the equilibrium behavior of a group is fully characterized by its size and that equilibrium strategies are of the threshold type. Specifically, we find that all groups whose size is smaller than a given threshold outbreed, while larger groups inbreed. These results hold for large enough  $n$ , and rest on the main axioms that we discussed in the previous section.

We start with the one-sided model, where the arguments are simpler. Next, we address the two-sided model where the same idea is essentially at work, although some details differ.

**PROPOSITION 3 (Threshold Equilibrium - One-Sided Model)** *Consider the one-sided variant of the search model, with preferences satisfying (2)-(3). Assume that either axiom SM holds or*

that WM holds and the utility function  $U$  is linear. Also assume that the outbreeding cost is low in the following sense:

$$[1 - P_{\eta, \infty}(0)] V(1) > c. \quad (7)$$

Then, there exists some  $\hat{n}$  and a unique (finite)  $\tau^* \geq 2$  such that if  $n \geq \hat{n}$ , the equilibrium strategy profile  $\gamma^* = (\gamma_l^*)_{l=1}^q$  satisfies:

$$\gamma_l^* = I \Leftrightarrow n_l \geq \tau^* \quad (l = 1, \dots, q). \quad (8)$$

The intuition behind this result is easy to grasp. In view of the monotonicity axiom, smaller groups experience a larger advantage in searching over the whole population rather than only within the group. This advantage will outweigh the cost of outbreeding only for groups up to a given threshold size, which are the groups that outbreed in equilibrium.

One may worry that, in many cases, the threshold  $\tau^*$  established by Proposition 3 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 is low enough, the equilibrium threshold can be made arbitrarily large if  $c$  is low enough. For completeness, we state this conclusion in the following corollary:

**COROLLARY 1** *Under the assumptions made in Proposition 3, for any  $\tau_0$  there is some  $\bar{c} > 0$  such that if  $c < \bar{c}$  then the equilibrium threshold  $\tau^* \geq \tau_0$ .*

The same ideas underlying the one-sided case holds for the two-sided model, only that then the advantage of outbreeding depends (endogenously) on the size of the outbreeders' pool. The following proposition states that as long as small groups make up for a large enough (that is, non negligible) share of the whole population, the threshold result carries over unaffected to this case.

**PROPOSITION 4 (Threshold Equilibrium - Two-Sided Model)** *Consider the two-sided variant of the search model, with preferences satisfying (2)-(3). Assume that either axiom SM holds or that WM holds and the utility function  $U$  is linear. Also assume that the outbreeding cost  $c$  satisfies condition (7). Then, there exists some  $\hat{n}$ ,  $\alpha > 0$ , and a unique (finite)  $\tau^* \geq 2$  such that if  $n \geq \hat{n}$  and  $\sum_{l:n_l < \tau^*} n_l > \alpha n$ , the equilibrium strategy profile  $\gamma^* = (\gamma_l^*)_{l=1}^q$  satisfies:*

$$\gamma_l^* = I \Leftrightarrow n_l \geq \tau^* \quad (l = 1, \dots, q).$$

Note that in the two-sided model the strategy profile in which all groups outbreed is always a trivial equilibrium of the game. In general, there may be multiple equilibria, associated to different thresholds  $\tau$  that lead to different numbers of outbreeding groups. Among these, the threshold  $\tau^*$  in Proposition 3 is the largest one possible. Equilibria defined by a threshold  $\tau < \tau^*$  embody some manifestation of the coordination failure that, in its extreme version, is displayed by the trivial full-inbreeding equilibrium ( $\tau = 1$ ). Finally, it should be clear that the counterpart of Corollary 1 also

applies to the two-sided context, and the corresponding threshold  $\tau^*$  also increases unboundedly as the outbreeding cost becomes small.

## 4 Homophily: theory and evidence

As we shall see, the equilibrium inbreeding/outbreeding choice derived from the search model has strong implications for the type-composition of matches. In this section, we review some of these implications and compare them with empirical evidence. Our discussion is structured as follows. First, in Subsection 4.1, we introduce the measures that capture 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. We do this in the following two subsections: intra-group ties in Subsection 4.3 and inter-group ties in Subsection 4.4. Subsection 4.5 finally discusses additional empirical evidence on High School friendships.

### 4.1 Measuring homophily

To start with, a simple index of the "bias" in agents' matches is obtained by comparing the resulting mix of types with the theoretical mix implied by uniform random assortment. Specifically, for each group  $l$ , let us denote by  $m_{ll'}$  the average 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 average number of *total* matches of agents of type  $l$ . The ratio  $\frac{m_{ll'}}{m_l}$  measures the representation of type  $l'$  matches in the total matches of group  $l$ . Denoting by  $w_{l'}$  the relative population share  $\frac{n_{l'}}{n}$  of group  $l'$ , the difference

$$\Delta_{ll'} \equiv \left( \frac{m_{ll'}}{m_l} - w_{l'} \right) \quad (l, l' = 1, 2, \dots, q) \quad (9)$$

measures the bias towards group type  $l'$  in the matches of group  $l$ , and will be called the *excess representation* of type  $l'$  in the matches of  $l$ . As an important particular case, if we make  $l' = l$ , the ratio  $\frac{m_{ll}}{m_l}$  of same-type matches for agents of group  $l$  can be regarded as an index of homophily for this group and we denote it by  $H_l$ . We then refer to the difference  $H_l - w_l > 0$  as *excess homophily* for group  $l$ .<sup>12</sup> If prevailing ties were 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 that its maximal potential value,  $1 - w_l$ , is small to begin with. The index proposed by Coleman (1956), and recently employed in various papers (see Currarini, Jackson and Pin (2009, 2010a), Bramoullé

---

<sup>12</sup>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 search model.

and Rogers (2009)) addresses the problem by normalizing the excess homophily of group  $l$  by its maximal value  $1 - w_l$ :

$$C_l = \frac{H_l - w_l}{1 - w_l}. \quad (10)$$

This is the Homophily Coleman Index on which we shall base much of our subsequent analysis of the model.

## 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, 2010a) in their economic model of friendship.<sup>13</sup> 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,<sup>14</sup> 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.<sup>15</sup> 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 "marriage market." However, if a state has a city with more than 500.000 people, we split it in two:

---

<sup>13</sup>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).

<sup>14</sup>Racial groups are Whites, Blacks Hispanic and Asians.

<sup>15</sup>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."

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

In this section we first summarize the main empirical regularities in the patterns of in-group matches for both friendships and marriages, and then discuss whether this evidence can be traced to elements and theoretical implications of our model. In all pictures 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.

#### 4.3.1 Homophily Index

**Empirical Evidence** We start with the basic index of homophily measured by the fraction of the total number of ties of a group with individuals of that same group. We report in figure 1 the empirical pattern of this index with respect to groups' population shares for both friendships (left) and marriages (right). For the case of marriages, the main qualitative pattern is a sharp structural change in the relationship around a 10% level of population share. Before this level we identify a positive relation with significant<sup>16</sup> coefficient of about 5 (with zero intercept); beyond this level, we identify a coefficient of 0.4 (and a positive and significant intercept) for groups making up for more than 10% of total population. A similar, although less sharp, pattern applies to friendships, with a coefficient of about 3.4 for groups below 20% (and an intercept not different from zero) of population and of about 0.4 for larger groups (with significantly positive intercept).

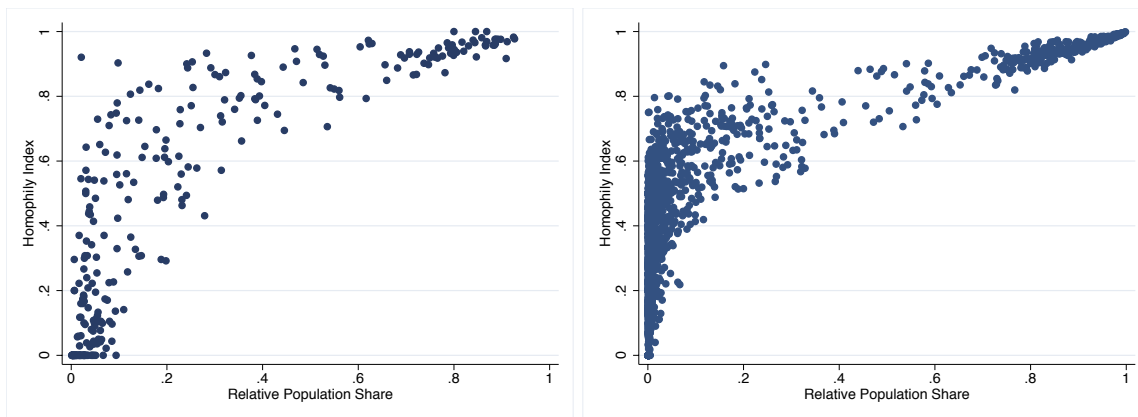
**Theory** This sharp structural change can be explained in terms of the qualitative switch in search behavior, from outbreeding to inbreeding, at the equilibrium threshold size. This threshold is uniquely determined in the one-sided variant of the model, and its value also defines the maximal threshold of the two-sided variant. Here, however, multiple equilibria, characterized by different thresholds, are possible.<sup>17</sup> This implies an homophily index of one for large inbreeding groups in both variants of the model. For small outbreeding groups, the exact theoretical predictions depend on what one assumes on the probability distribution governing the outbreeding search process. If this distribution is uniform (as in the case, for example, of the first illustration in Subsection 2.2), then outbreeding groups have an homophily index which equates their population shares in the

---

<sup>16</sup>When not otherwise specified, we mean a 99% significance level.

<sup>17</sup>This threshold size was defined in our propositions 3 and 4 in terms of groups absolute size. When aggregating across populations of different size one has to be careful about the difference between group size and population share, since groups of a given same population share but belonging to populations of different sizes may end up adopting different inbreeding/outbreeding strategies in equilibrium. This (together with other unmodeled heterogeneity - such as different costs of outbreeding) may be responsible for some of the variability in the picture, but would leave the qualitative features of the observed trends unaffected.

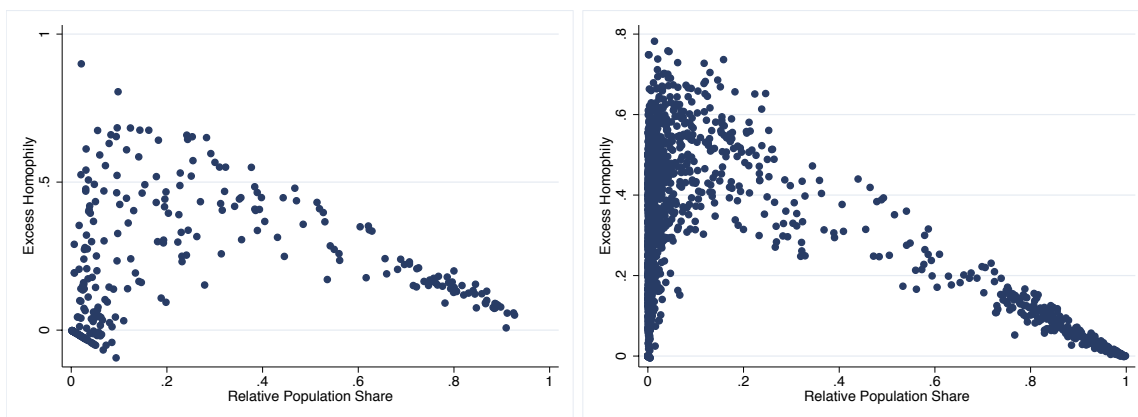
one-sided model, and an index that exceeds their population shares in the two-sided variant (this because in this case search is uniform on the restricted pool of outbreeders).



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

### 4.3.2 Excess Homophily

**Empirical Evidence** Figure 2 illustrates the *excess homophily* of groups as a function of their respective relative sizes, for both friendships (left panel) and marriages (right panel). This is simply obtained from 1 as the difference between each dot and the 45 degree line. Figure 2 highlights two important patterns: first, the excess homophily is increasing (and steeply so) for groups covering less than 10% of total population, and decreases linearly thereafter. Again, this is particularly clear in the case of marriages, where the number of observation allows for a lower variability.



2: Excess Homophily: High School Friendships (left) and Marriages (right).

**Theory** To account for the evidence of figure 2, let us first look at the two-sided variant of our model. Here, under the assumption that the distribution governing the outbreeders' search is uniform, small groups have a positive and increasing excess homophily. To see this, note that small groups find same-type matches with probabilities equal to their population share in the pool of outbreeders. Therefore, the excess homophily of outbreeders is given by  $(\frac{n_l}{n_O} - \frac{n_l}{n})$ , which is indeed increasing in  $n_l$  (and very steeply so, if  $n_O$  is small relative to  $n$ ).

Turning now to the one-sided model, our theory predicts a null excess homophily for all outbreeders, failing therefore to deliver the increasing part of figure 2. To account for this evidence (and for other regularities that will be discussed below) we enrich the search model with small frictions, that can be interpreted as either small mistakes made by agents in allocating their search time, as exogenous factors that interfere with agents' search or, more generally, as the mere effect of some residual randomness in social encounters.

Specifically, let us define by  $r_I$  some additional time/effort that each agent exerts within her own group no matter whether she has decided to inbreed or outbreed.<sup>18</sup> In particular,  $r_I$  can be thought of some residual "inbreeding" search that all agents perform due to exogenous constraints (think, for instance, of kinship when groups are defined by race).

The next proposition shows that such small search friction naturally generates the increasing excess homophily of small groups found in figure 2.

**PROPOSITION 5** *Consider any two outbreeding groups  $l$  and  $l'$  with  $w_l < w_{l'}$ . There exist some  $\hat{n}$  such that if  $n \geq \hat{n}$ , then  $H'_l - w'_l > H_l - w_l$ .*

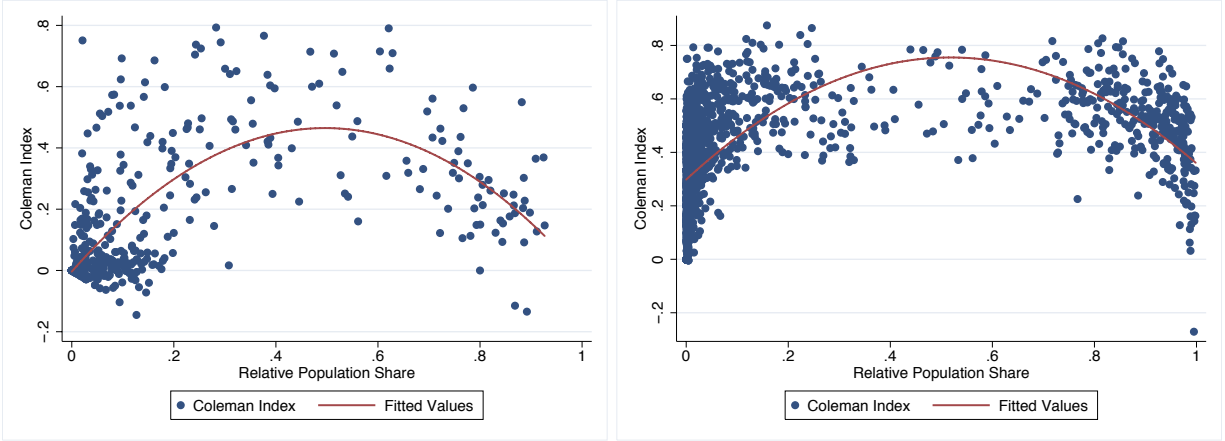
### 4.3.3 Coleman Index

**Empirical Evidence** We next turn to the empirical pattern of the Coleman Index, formally defined in (10). Figure 3 illustrates the relation between this index and the relative size of ethnic groups for both high school friendships (left panel) and marriages (right panel). The non linear and non monotonic pattern of the left panel was first identified for friendships in CJP; Figure 3 shows that this trend also characterizes U.S. marriages.<sup>19</sup> In particular, the Coleman index takes maximal values for middles sized groups and lower values for very small and very large groups.<sup>20</sup>

<sup>18</sup>If we think of search effort in terms of time allocation, then  $r_I$  and  $r_O$  can be thought as some (small) fraction of time during which social encounters happen independently of the agents' intentions.

<sup>19</sup>The fitted lines in figure 3 are obtained by regressing the index  $C_q$  on population shares and on the square of population shares. Details of the regressions are as follows ( $t$ -statistics are in brackets): for friendships:  $C = .4 + 2.1w - 2.2w^2$ ; for marriages:  $C = 0.29 + 0.76w - 0.70w^2$ . In both regressions, both coefficients for  $w$  and  $w^2$  are statistically significant at 99% level; the constant term is significant only for marriages.

<sup>20</sup>The main difference between the right and the left panels of figure 3 is that in the right panel 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.



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

**Theory** While the increasing branch of the parabola is a direct consequence of the increasing pattern of excess homophily (see the discussion in the previous sections and the proof of proposition 7 below), the decreasing part, finally approaching zero for large groups in the case of friendships, imposes further restrictions on the model. In fact, our model predicts a Coleman Index equal to  $\frac{1-w_l}{1-w_l}$  for large inbreeding groups, which is not well defined for  $w_l \rightarrow 1$ . predictions on the Coleman Index of large inbreeding groups are indeterminate in the limit for  $w \rightarrow 1$ . To account for the full range of the empirical evidence, we extend the assumption of search friction also to the inbreeding technology. We denote by  $r_O$  a small search intensity that all groups always use to search for agents outside their own group. This small friction can again be interpreted as some additional randomness, or probability of mistakes, that affects agents beyond their inbreeding/outbreeding choices.

The following propositions show that, if frictions are small compared to total search intensity, and if total population is large, the threshold equilibrium induced by our model generates the full non monotonic pattern observed in figure 3. To be sure, also note that very small frictions do not affect in any significant way the linear and decreasing behavior of excess homophily for inbreeding groups discussed before. Thus, in the end, such a friction-laden enrichment of the model allows us to account for the empirical evidence displayed by the Coleman Index in Figure 3, without impairing its ability to account for the pattern on excess homophily depicted in Figure 1.

The ensuing set of results provides a range of conclusions that, in combination, map a behavior for the Coleman Index that is qualitatively in line with that observed in Figure 3. Succinctly, they show that this index starts at very low levels and rises with size for very small groups, becomes close to 1 for intermediate-sized groups, and then decreases with size, finally becoming very small (and negative) as relative size approaches 1.

For notational convenience, denote by  $m(\eta, \theta) \equiv \mathbb{E}[\nu(\eta, \theta)]$  the expected number of potential matches<sup>21</sup> when search effort is  $\eta$  and the pool size is  $\theta$ . The first proposition implies that very small groups, and in particular outbreeding groups, are characterized by a very small Coleman Index if  $r_I$  is small.

**PROPOSITION 6** *Consider any group  $l$  such that  $n_l < \tau$ , where  $\tau$  is as specified in Proposition 3. Then, there exists some  $\hat{n}$  such that if  $n \geq \hat{n}$ , then  $C_l$  is bounded above by the term  $\frac{m(r_I, n_l)}{m(\eta+r_O, \infty)}$ .*

This is a good time to discuss our continuity axiom CON, which plays a significant role in Proposition 6. This axiom ensures that as the relative size of a group becomes of negligible measure, the probability of finding any of its members through search in the large pool (that is, through outbreeding) becomes zero. The use of this axiom, here and in the following propositions, implies that outbreeding search by members of small groups does not contain any bias in favor of one's own group.<sup>22</sup> In our model, therefore, for the sake of clarity, any such bias is fully exerted through the choice of inbreeding.

The next result, pertaining again to outbreeding groups shows that for these groups the Coleman Index is ordered according to size.

**PROPOSITION 7** *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  $C_l \leq C_{l'}$ .*

The next result pertains to 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. These groups, as we now establish, are characterized by arbitrarily high levels of the Coleman Index for low levels of search frictions.

**PROPOSITION 8** *Given any  $\varepsilon > 0$ , there exist some  $\delta_1 > 0$ ,  $\delta_2 > 0$ ,  $\delta_3 > 0$  and  $\hat{n}$  such that if  $n \geq \hat{n}$  and  $\frac{m(r_O, \infty)}{m(\eta+r_I, \infty)} \leq \delta_3$  then any group  $l$  with relative size  $\delta_1 > \frac{n_l}{n} > \delta_2$  has  $C_l \geq 1 - \varepsilon$ .*

Finally, the next two results complete the present analysis by specifying how the homophily index changes with group size among relatively large groups (that inbreed). First, Proposition 9 shows that, among groups that inbreed and have a nonnegligible relative size, the Coleman index decreases as size grows larger. Second, Proposition 10 establishes 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.

<sup>21</sup>In expected terms, successful matches are proportional to potential ones.

<sup>22</sup>Note also that the CON axiom is the only assumption we make on the type composition of matches induced by outbreeding search.

**PROPOSITION 9** Consider any two groups,  $l$  and  $l'$ , such that their relative sizes are bounded away from 1 and such that  $\frac{n_{l'}}{n} - \delta_1 \geq \frac{n_l}{n} > \delta_2$  for some  $\delta_1, \delta_2 > 0$ . Then, there exists some  $\hat{n}$  such that if  $n \geq \hat{n}$ ,  $C_l < C_{l'}$ .

**PROPOSITION 10** There exist some  $\delta_2 > 0$  and  $\hat{n}$  such that if  $n \geq \hat{n}$ , then any group  $l$  with relative size  $\frac{n_l}{n} \geq 1 - \delta_2$  has  $C_l < 0$ .

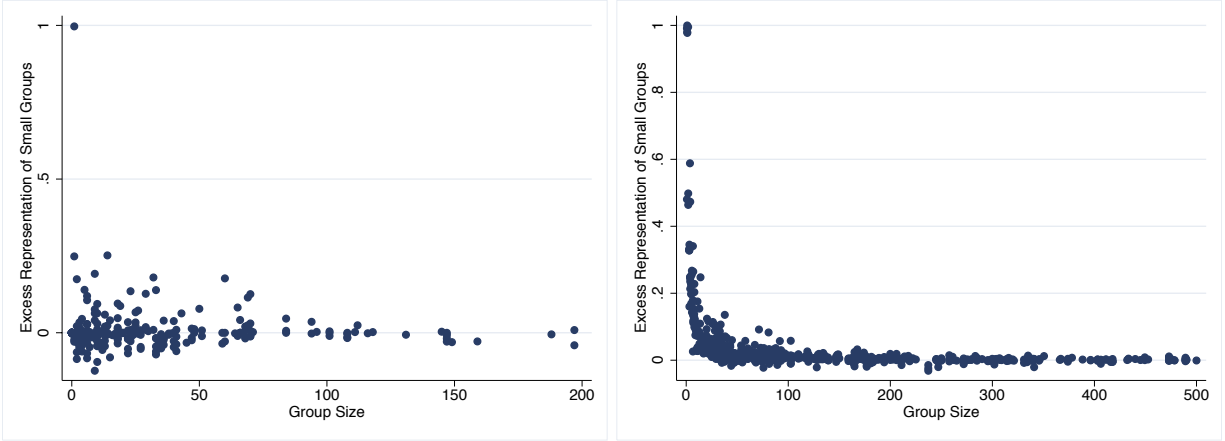
#### 4.4 Cross-group ties

In this section we look at cross-group ties. In particular, we are interested in the empirical distribution of such ties across groups of different size, and in how this may be driven by some of the properties of our search equilibrium. As explained at the outset of Subsection 4.1, we decompose the total number of ties of every group  $l$  by the group  $l'$  of their destination. So, for instance, in the case of friendship ties, we compute the proportion of Black friends, or Hispanic friends, or Asian friends among the total number of friends enjoyed by White individuals. This allows us to compute the *excess representation*  $\Delta_{ll'}$  across various racial groups  $l$  and  $l'$ , as defined in (9).

The predictions of our search model for cross-group ties crucially depend on which variant one adopts. In the one-sided variant, outbreeders search in the whole pool, and, if search is uniform, find each group at rates that follow the relative sizes of these groups. In contrast, the two-sided variant predicts that outbreeders search in the restricted pool of outbreeders. Thus, if search is uniform, the matches of outbreeding groups should display an excess representation of other outbreeding groups. 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 excessive representation of small groups in friendships and marriages. In figure 4, 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. In the figure we report on the specific thresholds of  $x = 80$  for friendships and  $x = 100$  for marriages. These cases are shown for illustrative purpose, and qualitatively similar pictures obtain when we fix different small thresholds.

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 model, the critical size suggested by these pictures is marked by 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.



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

#### 4.5 More on Friendships: School Size, Homophily and Popularity

The cornerstone of our theory is the assumption that the size of small groups limits the search opportunities of their members, and that this constraint is relaxed as the size of a group increases. As we have seen, this assumption has implied a difference in the equilibrium behavior of small and large groups. In this final section we show that the effect of group size can help explain other empirical regularities on homophily in American high schools.

##### 4.5.1 School size and homophily

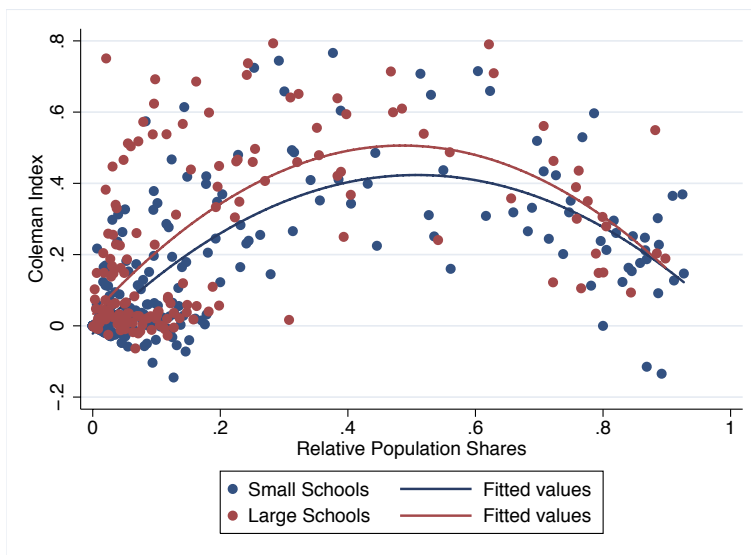
School size (in terms of total number of students) has been shown by Currarini, Jackson and Pin (2010a) to significantly affect the homophilous bias in the friendships of students of all ethnic groups. In particular, larger schools (with more than 1000 students) are shown to display uniformly larger 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, and to mostly reflect a difference in the meeting opportunities that students face in small versus large schools. This 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.<sup>23</sup>

Consider again the equilibrium threshold  $\tau$ , below which a group finds it profitable to outbreed. As it is shown in Proposition 8, 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

<sup>23</sup>We are grateful to Matt Jackson for pointing out to us this property of our model.

in terms of the relative size of groups (that is, their fraction of the total population), which is measured on the x-axis of Figure 5 and of all previous figures in the present paper. 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  $\tau$  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  $\tau$ . 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.

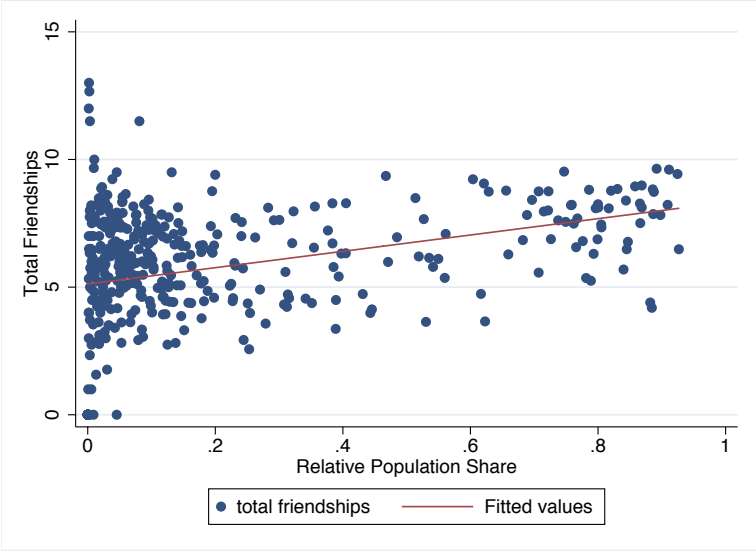


5: Coleman Index in Friendships: Small Schools (< 1000) vs. Large Schools (> 1000).

#### 4.5.2 Search, Group size and Popularity

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 search intensity, so that larger groups, facing better prospects in terms of type-mix of friends, search more intensely. This aspect is absent from our model, where search intensity is exogenous and constant across groups.



6: Total Friendships by Relative Population Share.

In the context of our model, group size can still affect the total number of matches if some (small) groups suffer from search inefficiency in equilibrium. This is always the case in the presence of the search friction  $r_I$ , which "forces" members of all groups to direct some search effort towards their own group only. Being exerted in pools of different sizes, this part of search is in fact subject to our monotonicity axioms, and 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 constraints that a small size imposes on search. This effect vanishes for larger groups, for which the effect of relative population shares takes over.

## 5 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.

Naturally, the model is too schematic to account for some of interesting dimensions involved in the phenomenon of homophily in social contexts. In particular, an important issue by-passed by our approach is the role of preferences. In CJP (Currarini, Jackson and Pin (2009)) preferences biased in favor of own type are key to understanding the differences in total friendships observed by small and large groups. Introducing biased preferences for own type in the present context would affect the incentives to outbreed by small groups, which would have lower incentives to outbreed due to a worsening the mix of the search pool, much along the same lines as in CJP. To introduce this aspect into the model would enrich the incentives structure of our model, and seems an interesting avenue for further work.

The purpose of our paper, however, was not to construct a comprehensive model of homophily but to propose a microfoundation of the meeting bias that, despite its simplicity, is consistent with

many of the observed empirical patterns. Yet, 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 an important issue for future theoretical and empirical research.

## References

- [1] Allport, W. G. (1954): *The Nature of Prejudice*. Cambridge, MA: Addison-Wesley.
- [2] Barr, A., Dekker, M. and M. Fafchamps (2009) “Bridging the gender divide: An experimental analysis of group formation in African villages”, ASC Working Paper 87 / 2009.
- [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] Buhai, S. and M. van der Leij (2009) “A Social Network Analysis of Occupational Segregation”, mimeo.
- [6] Coleman, J. (1958): “Relational Analysis: The Study of Social Organizations With Survey Methods”, *Human Organization* 17, 28-36.
- [7] Currarini, S., M.O. Jackson, and P. Pin (2009) “An Economic Model of Friendship: Homophily, Minorities and Segregation,” *Econometrica* 77, No. 4, 1003–1045.
- [8] Currarini, S., M.O. Jackson, and P. Pin (2010a) “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.
- [9] Currarini, S., M.O. Jackson, and P. Pin (2010b) “Long Run Integration in Social Networks”, mimeo.
- [10] Dixit, A. (2003), “Trade Expansion and Contract Enforcement”, *Journal of Political Economy* 111(6), 1293- 1317.

- [11] Franz, S., M. Marsili and P. Pin (2008), “Observed choices and underlying opportunities”, mimeo.
- [12] Giles, M. W. (1978), “White Enrollment Stability and School Desegregation: A Two- Level Analysis” *American Sociological Review* 43, 2448-64.
- [13] Golub, B. and M. O. Jackson (2010), “How Homophily Affects the Speed of Contagion, Best Response and Learning Dynamics ”, arxiv version: <http://arxiv.org/abs/0811.4013> physics.soc-ph.
- [14] 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.
- [15] Marsden, P.V. (1987): “Core Discussion Networks of Americans,” *American Sociological Review* 52, 122-313.
- [16] Marsden, P.V. (1988): “Homogeneity in Confiding Relations,” *Social Networks* 10, 57-76.
- [17] McPherson, M., L. Smith-Lovin and J. M. Cook (2001): “Birds of a Feather: Homophily in Social Networks”, *Annual Review Sociology* 27, 415-44.
- [18] Moody, J. (2001): “Race, School Integration, and Friendship Segregation in America” *The American Journal of Sociology*, 107(3), 679-716.
- [19] Patacchini, E. and Y. Zenou (2008) “Ethnic Networks and Employment Outcomes”, IZA Discussion Papers 3331.
- [20] Stajc, W. (1990), “The Collector’s Problem with Group Drawings ”, *Advances in Applied Probability*, 22(4), 866-882.
- [21] Vega-Redondo, F. (2007): *Complex Social Networks*, Econometric Society Monograph Series, Cambridge: Cambridge University Press.

## Appendix

As indicated in the text, we start by showing that the optimal allocation of intensity between inbreeding and outbreeding must occur at extreme values if the population is large and the following axiom holds:

**ADD (Additivity in Search)** For all  $(\eta_1, \eta_2)$  such that  $\eta_1 + \eta_2 = \eta$ , let  $\hat{\nu}(\eta_1, \eta_2, \infty) \equiv \nu(\eta_1, \infty) + \nu(\eta_2, \infty)$  be the random variable associated with the sum of the outcomes of two stochastically independent search processes with intensities  $\eta_1$  and  $\eta_2$  on an infinite pool. Then  $\nu(\eta_1, \eta_2, \infty) = \nu(\eta, \infty)$ .

**LEMMA 1** *Assume that the random variable  $\nu$  satisfies axiom ADD, and either SM holds or WM holds and  $U$  is linear. Consider any agent belonging to a group of size  $n_l$  and let  $\eta_I$  and  $\eta_O$  denote the intensity she devotes to inbreeding and outbreeding, respectively. Then, if  $n$  is large enough, the optimal choice of each agent either involves  $\eta_I = \eta$  (and  $\eta_O = 0$ ) or  $\eta_O = \eta$  (and  $\eta_I = 0$ ).*

**Proof** Assume first that SM holds and consider any  $\eta_I$  and  $\eta_O$  such that  $\eta_I + \eta_O = \eta$ . If  $n$  is large enough, the random variable determining the number of distinct draws can be approximated by the sum of independent random variables  $\nu(\eta_1, n_l) + \nu(\eta_2, \infty)$ , which in turn is FSD by  $\nu(\eta_1, \infty) + \nu(\eta_2, \infty)$ , equal to  $\nu(\eta, \infty)$ , by virtue of Axiom ADD. This implies that if  $\eta_O > 0$  and the outbreeding cost is to be paid, the optimal value is  $\eta_O = \eta$ . Otherwise, the optimal  $\eta_O = 0$ , which implies  $\eta_I = \eta$ . This proves the desired conclusion under SM. If only WM holds and  $U$  is linear, the argument is analogous.

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

**Proof of Proposition 1**

Let us denote by  $p(d, \eta, \theta)$  the probability of  $d$  distinct draws from  $\eta$  draws with replacement out of a pool of size  $\theta$ . As shown by Staje (1990):

$$p(d, \eta, \theta) = \binom{\theta}{d} \sum_{j=0}^d (-1)^j \binom{d}{j} \left( \frac{d-j}{\theta} \right)^\eta$$

Let us consider the ratio of  $p(d, \eta, \theta)$  to  $p(d, \eta, \theta + 1)$  :

$$\frac{p(d, \eta, \theta)}{p(d, \eta, \theta + 1)} = \frac{\binom{\theta}{d} \sum_{j=0}^d (-1)^j \binom{d}{j} \left( \frac{d-j}{\theta} \right)^\eta}{\binom{\theta+1}{d} \sum_{j=0}^d (-1)^j \binom{d}{j} \left( \frac{d-j}{\theta+1} \right)^\eta} \quad (11)$$

which can be written as:

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

which reduces to:

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

Note that for  $d = 1$  this yields:

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

Note also that for all admissible values of  $\theta$  and  $d$ , the ratio  $\frac{p(d, \eta, \theta)}{p(d, \eta, \theta + 1)}$  is decreasing in  $d$ . Since these are probability distributions, we conclude that there exists  $\bar{d}$  such that  $\frac{p(d, \eta, \theta)}{p(d, \eta, \theta + 1)} > 1$  for all  $d > \bar{d}$ . This implies that  $p(d, \eta, \theta + 1)$  First Order Stochastically Dominates  $p(d, \eta, \theta)$ .

### Proof of Proposition 2

We first derive the expected number of distinct draws that agent  $i$  obtains from the set  $N \setminus L$  by means of  $\eta$  independent draws with replacement out of the set  $N$ , for any given subset  $L \subset N$  of cardinality  $l$ . The set  $L$  should be interpreted here 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  $i$ ). This expected number is given by:

$$(\theta - l) \cdot p(\eta, \theta) \quad (12)$$

where

$$p(\eta, \theta) = \left(1 - \left(\frac{\theta - 1}{\theta}\right)^\eta\right). \quad (13)$$

is the probability that an agent in a pool of size  $\theta$  is found by means of  $\eta$  draws with replacement from that pool (see Stadjé (1990)).

The expected value of the random variable  $\nu_{ap}(\eta, \theta)$  can now be obtained by averaging (12) over all possible values of  $l$  using the binomial distribution. We obtain the following:

$$E(\nu_{ap}(\eta, \theta)) = \sum_{l=0}^{\theta} \binom{\theta}{l} p(\eta, \theta)^l (1 - p(\eta, \theta))^{\theta - l} \cdot l + \sum_{l=0}^{\theta} \binom{\theta}{l} p(\eta, \theta)^l (1 - p(\eta, \theta))^{\theta - l} \cdot (\theta - l) \cdot p(\eta, \theta). \quad (14)$$

We can rewrite the second sum in (14) as follows by extracting the term  $\theta \cdot p(\eta, \theta)$  which does not depend on  $l$ :

$$\theta \cdot p(\eta, \theta) \sum_{l=0}^{\theta} \binom{\theta}{l} p(\eta, \theta)^l (1 - p(\eta, \theta))^{\theta - l} - p(\eta, \theta) \sum_{l=0}^{\theta} \binom{\theta}{l} p(\eta, \theta)^l (1 - p(\eta, \theta))^{\theta - l}. \quad (15)$$

Note that the second term of (15) is just  $\theta \cdot p(\eta, \theta)^2$ , while the first term is simply  $\theta \cdot p(\eta, \theta)$ . So we get:

$$E(m) = p(\eta, \theta) \cdot \theta + p(\eta, \theta) \cdot \theta - p(\eta, \theta)^2 \cdot \theta = f(\theta, \eta)$$

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  $\frac{\partial f(\theta, \eta)}{\partial \theta}$  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:

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

The above condition is a direct consequence of the strict concavity of the log function. In fact, strict concavity implies that  $2\eta$  times the increase of the log function from  $\theta - 1$  to  $\theta$  is more than the increase from  $\theta - 1$  to  $2\eta + \theta - 1$ .

**Proof of Proposition 3** First we note that in the one-sided model, the payoff of a player  $i$  of an outbreeding group  $l$  of finite size in a group-symmetric profile  $\gamma$  is independent of the choice of groups other than  $l$ . The expected payoff  $\pi_i(\gamma)$  for an individual  $i$  of an outbreeding group  $l$  is given by the expression:

$$\pi_O(n_l) = \sum_{\nu_i=0}^n P_{\eta, \infty}(\nu_i) V(\nu_i) - c - \delta(n), \quad (16)$$

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

$$\pi_I(n_l) = \sum_{\nu_i=0}^n P_{\eta, n_l}(\nu_i) V(\nu_i). \quad (17)$$

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

$$[1 - P_{\eta, \infty}(0)] V(1) > c.$$

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

Recall now that  $U(y_i + 1) \geq U(y_i)$  for all  $y_i \geq 1$  and  $U(1) > U(0)$ . Therefore, either the SM axiom or the WM axiom combined with the linearity of  $U$  imply that, for all  $n_l$ ,

$$\pi_I(n_l + 1) - \pi_I(n_l) > 0. \quad (18)$$

Let now  $\tau$  be the lowest integer  $n_l$  such that

$$\pi_I(\tau) \geq \sum_{\nu_i=0}^n P(\nu_i(\eta, \infty)) V(\nu_i) - c. \quad (19)$$

Then, both if (19) 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 the desired conclusion.

**Proof of Proposition 4** When  $n$  grows large, the random variable  $\nu_i(\eta, \alpha n)$  approaches the limit  $\nu_i(\eta, \infty)$ . This implies that (16) again has its term  $\delta(n)$  vanish as  $n \rightarrow \infty$ . Then, one can proceed as in Proposition 3 to prove that the same threshold  $\tau^*$  found in that proposition applies in the present case.

**Proof of Proposition 5** For simplicity, consider two outbreeding groups  $l$  and  $l'$  whose cardinality differ in just one individual, i.e.  $n_{l'} = n_l + 1$ , and let  $\Delta \equiv (H_{l'} - w_{l'}) - (H_l - w_l)$  denote the difference in the Excess Homophily of the two groups. We need to show that  $\Delta > 0$ . By the SM axiom (or its weaker version WM),  $m(r_I, n_{l'}) - m(r_I, n_l) > 0$  for all  $r_I$ . Let us write:

$$\Delta = \left[ \frac{m(r_I, n_{l'})}{m(r_I, n_{l'}) + m(r_O + \eta, \infty)} - \frac{n_{l'}}{n} \right] - \left[ \frac{m(r_I, n_l)}{m(r_I, n_l) + m(r_O + \eta, \infty)} - \frac{n_l}{n} \right] \quad (20)$$

$$= \frac{m(r_I, n_{l'})}{m(r_I, n_{l'}) + m(r_O + \eta, \infty)} - \frac{m(r_I, n_l)}{m(r_I, n_l) + m(r_O + \eta, \infty)} - \frac{1}{n}. \quad (21)$$

Since  $m(r_I, n_{l'}) - m(r_I, n_l) > 0$ , the difference

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

is strictly positive, and is uniformly bounded away from zero. It is clear that for  $n$  large enough,  $\Delta$  is strictly positive.

**Proof of Proposition 6** First note that, since  $\tau$  is independent of  $n$ , then  $\frac{n_l}{n} \searrow 0$  as  $n \nearrow \infty$ . Thus, for  $n$  large enough  $C_l$  can be approximated by the ratio  $H_l$ , given by  $(\frac{m_{ll}}{m_l} - w_l)$ .

Let us now consider the expected number of same type matches  $m_{ll}$  for group  $l$ . By our continuity axiom CON, if group  $l$  outbreeds and  $\frac{n_l}{n} \searrow 0$  for large enough  $n$ , we can approximate the expected value of  $m_{ll}$  by  $m(r_I, n_l)$ , and that of  $m_l$  by  $[m(r_I, n_l) + m(\eta + r_O, \infty)]$ . Then we obtain:

$$C_l = \frac{m(r_I, n_l)}{m(r_I, n_l) + m(\eta + r_O, \infty)}.$$

**Proof of Proposition 7** For simplicity, consider two outbreeding groups  $l$  and  $l'$  whose cardinality differ in just one individual, i.e.  $n_{l'} = n_l + 1$ , and let  $\Delta \equiv C_{l'} - C_l$  denote the change in the Coleman index. Since passing from  $l$  to  $l'$  the denominator of the Coleman Index decreases, in order to establish the desired conclusion (i.e. that  $\Delta > 0$ ) it is enough to argue that the numerator increases. This is proved in Proposition 5

**Proof of Proposition 8** A preliminary observation is that, if  $n$  is large enough, then since  $\frac{n_l}{n}$  is bounded away from zero by  $\delta_2$  it must be that  $n_l \geq \tau$  (where  $\tau$  is as in Proposition 3) and

therefore group  $l$  inbreeds. Moreover, for large  $n$  its size can be so large that its ratio  $(\frac{m_l}{m_l} - w_l)$  can be approximated by  $\frac{m(\eta+r_I, \infty)}{m(\eta+r_I, \infty)+m(r_O, \infty)}$ . This in turn allows its index  $C_l$  to be approximated as follows:

$$C_l \simeq \frac{\frac{m(\eta+r_I, \infty)}{m(\eta+r_I, \infty)+m(r_O, \infty)} - \frac{n_l}{n}}{1 - \frac{n_l}{n}}.$$

An appropriate choices of  $\delta_3$  ensures that the term

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

is close enough to 1 and ensures the result.

**Proof of Proposition 9** Consider two groups,  $l$  and  $l'$ , whose relative sizes are bounded below by some positive number  $\delta_2$ . As  $n$  becomes large, both groups must exceed the threshold  $\tau$  specified in Proposition 3, so both find it optimal to inbreed. Then, by invoking the usual approximations of their corresponding Coleman Index (which again presume that  $n$  is large enough and  $\frac{n_{l'}}{n} - \frac{n_l}{n} \geq \delta_1$  for some positive  $\delta_1$ ), the desired conclusion reads:

$$\frac{\frac{m(\eta+r_I, \infty)}{m(\eta+r_I, \infty)+m(r_O, \infty)} - \frac{n_l}{n}}{1 - \frac{n_l}{n}} < \frac{\frac{m(\eta+r_I, \infty)}{m(\eta+r_I, \infty)+m(r_O, \infty)} - \frac{n_{l'}}{n}}{1 - \frac{n_{l'}}{n}}$$

or

$$\frac{n - n_l \frac{m(\eta+r_I, \infty)}{m(\eta+r_I, \infty)+m(r_O, \infty)}}{n - n_l} < \frac{n - n_{l'} \frac{m(\eta+r_I, \infty)}{m(\eta+r_I, \infty)+m(r_O, \infty)}}{n - n_{l'}}$$

which holds if, and only if,  $n_{l'} < n_l$ . The proof is thus complete.

**Proof of Proposition 10** Given  $\eta$ ,  $r_O$ , and  $r_I$ , choose  $\delta_2 < \frac{1}{2} \frac{m(r_O, \infty)}{m(\eta+r_I, \infty)+m(r_O, \infty)}$ . Then, it is straightforward to see that if  $1 - \frac{\delta_2}{2} \geq \frac{n_l}{n} \geq 1 - \delta_2$  then  $C_l$  - whose sign is the sign of the term  $\frac{m(\eta+r_I, \infty)}{m(\eta+r_I, \infty)+m(r_O, n-n_l)} - \frac{n_l}{n}$  for large  $n$  - is negative. To see this, note that  $m(r_O, \infty) > m(r_O, n - n_l)$  for all  $n_l$ , and that, if  $\delta_2$  is in the assumed range, then:

$$\frac{n_l}{n} \geq 1 - \delta_2 > \frac{2m(\eta + r_I, \infty) + m(r_O, n - n_l)}{2m(\eta + r_I, \infty) + 2m(r_O, n - n_l)} > \frac{m(\eta + r_I, \infty)}{m(\eta + r_I, \infty) + m(r_O, n - n_l)}.$$