

Homophily and Search

Sergio Currarini*

Fernando Vega Redondo[†]

April 26, 2010

Abstract

We study the formation of social ties among heterogeneous agents in a model where who meets who is determined through a random search process. Our aim is to understand how the incentives to connect with similar agents vary across social groups, and what role is played by groups' sizes. The key element of our approach is that search is more effective in large pools. We show that search equilibrium is characterized by a threshold in terms of group size, so that larger groups only search among similar agents, while smaller groups search among the whole population. Under the assumption that search is subject to small frictions, this type of equilibrium behaviour is shown to generate patterns of homophily which are consistent with empirical evidence from diverse social environments, such as high school friendships and interethnic marriages. In particular, homophily is largest in medium sized groups, and most inter-group ties occur between members of very small groups.

Keywords: Homophily, search, social networks, segregation.

JEL Classification: D7, D71, D85, Z13.

*Dipartimento di Scienze Economiche, Università di Venezia and School for Advanced Studies in Venice (SSAV).
Email: s.currarini@unive.it.

[†]European University Institute, Florence. Email: Fernando.Vega@eui.eu.

1 Introduction

A pervasive feature of social and economic networks is that contacts tend to be more frequent among similar agents rather than among dissimilar ones. This pattern, usually referred to as "homophily", applies to many types of social interaction, and along many dimensions of similarity.¹ The presence of homophily has important implications on how information and other aspects of social communication flow on the network of social contacts, and to what degree distance in characteristics translates into distance in the network. It is therefore important to understand more about the generative process of social networks, and how agents attitudes and 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 affects the distribution of intra- and inter-group links by affecting the meeting opportunities of agents.² If ties are formed uniformly at random, agents will end up meeting a fraction of members of a given group that reflect the group's population share. So, large groups, whose members are met with higher probabilities, will tend to be less "open" than smaller groups in terms of the fraction of ties made with dissimilar agents. Different groups are therefore characterized by different "baseline" degrees of homophily, even in the absence of biases in their attitudes towards dissimilar agents or in their meeting opportunities.

What is most striking about the empirical evidence of many social networks is that the observed homophily is often in excess of this baseline level, suggesting that the generative process of encounters is far from uniformly random, and is likely to be biased by agents' selective choices of assortment and/or by selective constraints that agents face in meeting each other. This has important implications for policy since agents' behavior may lead to actual segregation patterns that depart dramatically from the baseline level implied by the types' distribution (which can, as in the case of schools, be affected by policy makers).

In this paper we explore the patterns of homophily that result from a model where searching for social connections is more effective in larger pools. We introduce this assumption in the form of a general axiom on the search technology, and then provide several illustrations of search environments in which this axiom is justified. In these illustration, small groups suffer from redundancy in search (search is made through random draws with replacement or among friends of friends) or from lack of variety (agents have some taste for variety in the characteristics of their match and these characteristics are correlated with groups, as in Dixit 's (2003) model of trade).

Agents are partitioned in type homogeneous groups, and choose whether to direct their search only within their group ("inbreeding") or towards the whole population ("outbreeding"). Outbreeding implies a fixed cost (this may come from cultural, geographical, linguistic barriers to mixing

¹See McPearson, Smith-Lovin and Cook (2001) for an inclusive survey of the sociological literature on homophily.

²The notion that certain opportunities of encounters enhance the probability of tie formation is already present in Alport's (1954) contact theory

types), which provides immediate incentives to search only within one’s own group, and generates homophilous behaviour in the simplest and starkest manner.³ The crucial implication of our assumption on search effectiveness is that targeting the population at large has some efficiency advantage, that may counteract the cost of outbreeding. Most importantly, this advantage is largest for small groups’ members, which end up facing the strongest incentives to ”outbreed”.

We show that equilibrium strategies involve a threshold in group sizes, so that small groups outbreed, while larger groups inbreed. So, in equilibrium small groups’ members direct their search towards the whole populations, while larger groups only look inwards. We then focus on two equilibrium patterns of intra- and inter-group ties, of which we find evidence in two different instances of social networks in which race and ethnicity are significant determinants of assortative mixing: American high school friendships from the Add Health dataset and marriages from IPUMS census data.⁴

The first pattern characterizes the distribution of inter-group ties across groups of different sizes. If the formation of a tie requires mutual consent, then the threshold structure of equilibrium implies that most inter-group ties will occur between members of small groups. This because inter-group ties directed towards large groups’ are not reciprocated. Moreover, since small outbreeding groups uniformly search in the restricted pool of outbreeders, these small groups find each other at rates that exceed their population shares, and are found by large inbreeding groups at rates lower than populations shares.

The second pattern has to do with the share of intra-group ties that groups forms, and how this relates to groups’ sizes. From the above arguments, small outbreeding groups end up forming intra-group ties at rates that exceed population shares. On the other hand, large inbreeding groups only form intra-group ties. It follows that all groups exhibit homophily in excess of population shares. Evidence of this pervasive homophily was documented in recent papers by Currarini, Jackson and Pin (2009, 2010), where it is shown that the Coleman Homophily Index⁵ is positive for almost all groups, and follows a non monotonic trend, with small positive values for very small and very large groups, and largest values for middle sized groups. As we document in the present paper, this non monotonic trend is not specific to high school friendships, and we find a similar hump shaped curve in U.S. interethnic marriages. In Section 3 we show that the threshold equilibrium of our search game generates this non linear relation between the Coleman index and group size. This result is obtained by assuming that search is subject to small frictions, so that a ”small” number of different type friends may always be met independently of one’s inbreeding/outbreeding decision.

³Our purpose is here not to explain why agents are homophilous, but rather to explain some empirical evidence about the patterns of homophily across groups. The assumption of a fix cost to target different agents is therefore just a simple way to generate the type of behaviour of which we want investigate the main patterns.

⁴see section 3 for a detailed description of the employed datasets

⁵This index measures the difference between the the excess homophily of a group - that is, the difference between the share of ties with group members on total ties of the group and the population share of the group - normalized by the maximal potential excess homophily (see Section 3 for a detailed definition).

As we said, the equilibrium patterns of ties directly follow from the threshold property of chosen inbreeding/outbreeding strategies. Evidence of this "critical" size, after which a qualitative change in behaviour occurs, can be found in the sociological literature. The "tipping point" theory (see Giles (1978)) argues that race becomes relevant in the formation of social ties only after some minimal size has been reached by ethnic groups. Indeed, this size effect seems to underlie the observed non linear relation between school heterogeneity and friendship segregation found by Moody (2001) in the Add Health sample of American high schools. Segregation is there shown to stay nearly constant for small levels of racial heterogeneity, and then to sharply increase. If larger heterogeneity is associated with an increase in the size of minorities, the increase in segregation can be explained by a change in the inbreeding behavior of these minorities as they reach the critical size after which race becomes salient.

We finally wish to discuss the contribution of the present paper as compared to the quoted paper by Currarini, Jackson and Pin (2009) (CJP hereafter). There, the formation of ties is modeled as a search process, in which all agents randomly meet other agents from the same pool according to a uniform process. Homophily in excess of population shares is obtained for large groups, whose members search more intensively (for longer), and are therefore over-represented in the search pool. Since smaller groups end up being under-represented in the search pool, this is taken as evidence that a bias in the meeting process is responsible for the empirical evidence that all groups exhibit homophily in excess of population shares. A specific form of bias is then shown to generate the desired hump shaped relation between homophily and group size.

The meeting bias assumed in CJP should be viewed as a reduced form for unmodeled segregated opportunities, that may be due to the choice of agents to select where to look for friends (for instance, through race segregated clubs), or to physical or institutional constraints on meetings (e.g., academic tracking). The present model provides an analysis of the first of these potential sources of meeting bias, resulting from the inbreeding/outbreeding choices of agents. We do this in an extremely stylized model, with a minimal structure of search, in order to capture some basic aspects of homophily that underlie very different instances of social networks, such as friendships and marriages, whose micro details are likely to differ substantially.

The paper is organized as follows. Section 2 describes the search model, the main axioms and illustrations, and characterizes the equilibria of the game. Section 3 studies the implication of the model for homophily. It first describes the empirical evidence from friendships networks in U.S. high schools and from U.S. marriages. It then discusses the implication of the search equilibrium for homophily, showing that the obtained characterization fits empirical evidence. Section 4 concludes the paper.

2 The Search Model

We consider a set N of n agents. The set N is partitioned into groups, defined by a specific common trait (ethnic, linguistic, religious, etc.), that we call "type". Groups are indexed by q , and we denote by n_q the size of group q . Each agent i devotes η units of time (or effort) to search agents in N in order to find matches. The sole decision that agents take is how to allocate time between search among agents of their own group only and search on the whole population. 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).

The inbreeding/outbreeding decisions taken by all agents define the pools of other agents among which they search, and these pools in turn determine payoffs by shaping the matching outcomes. In what follows, for presentational convenience, we address these considerations in reverse order: first we describe how the pool size affects the number of matchings, secondly we discuss how matching probabilities affect payoffs and preferences, and thirdly we explain how agents' in-/out-breeding decisions shape their respective search pool. Finally, we shall combine all these to provide a formal specification of the search game that underlies our analysis.

2.1 Search pool and matching success

The central assumption that will underlie the analysis is a monotonicity assumption on how the number of matches depend on the size of the pool θ , by this meaning the set of agents that are involved in the agent's search activity (as we will explain, either as the target of the agent's search for potential matchings, or both as target and as source of potential matchings). Let us then define a "search profile" as a pair (η, θ) quantifying the amount of time devoted to search and the pool size. With each search profile (η, θ) we associate the random variable $\nu(\eta, \theta)$ expressing the resulting number of matches; the variable $\nu(\eta, \theta)$ takes values in $\{0, 1, \dots, n\}$.

Our key assumption that search is more effective in a larger pool is here embodied in the following two *axioms*.

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

This first axiom is 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. When First-Order Stochastic Dominance (FOSD) is too demanding (see the illustrations in the next section), we use a weaker axiom which embodies the same principle of size monotonicity without requiring as much as first order stochastic dominance.

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 $\mathbb{E}[\nu(\eta, \infty)]$ is well defined for all finite positive η .

2.2 Illustrations of the monotonicity axioms

We present three illustrations of search environments where the postulated monotonicity axioms hold. In the first two models, search is more effective in larger pools because size has the effect of reducing redundancies. The first model illustrates the key idea of redundancy in search through a stylized model of random draws with replacement. The second model shows how the same features arise when search of new social ties is made using existing ties as intermediaries. The third model 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 of size n makes a given number $\eta > 1$ of independent draws with replacement out of a population of size $\theta < n$. Search is of the "one-way flow" type, and matches for this agents can only come from his/her own draws (and not, for instance, if this agent is found by other agents through their own draws). This may represent information acquisition on the internet, where an agent only gets benefits from the sources of information he/she finds, or traditional marriage markets, in which one side (usually males) choose and propose to the other side, which is passive in the search process (in this example it is natural to have $\theta = n/2$).

Here, the matches that are relevant for an agent are the distinct draws that occur as an outcome of his η draws with replacement. After all, finding the same piece of information a second time does not add anything to what I already knew. Let us denote by $\nu_a(\eta, \theta)$ the associated random variable.

PROPOSITION 1 *The random variable $\nu_a(\eta, \theta)$ satisfies axiom SM. In particular, when the pool size grows large 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 over and over again is lower, and so the number of distinct draws increases (in a stochastic sense). While this result is obtained for a search process in which only one's own "active" draws generate matches, one can easily envisage search situations in which also "passive" draws (that is, other agents' draws through which an agent is found) matter in determining potential matches. Think, for instance, of friendship formation or of marriages in modern societies, where it is actually difficult to even distinguish the direction of search in a meeting. In these settings the above analysis

needs to be enriched to consider the effect of the enlargement of pool size on passive draws. Consider then the following two-way flow search process. For each given element i of a set N , let i make η random draws with replacement out of the set $N \setminus \{i\}$; at the same time, let each element j of the set $N \setminus \{i\}$ make η random draws with replacement out of a set $N \setminus \{j\}$, to which i belongs. Define then the 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 (note that here the variable θ is an index of group size, since agent i searches among $\theta - 1$ agents, and $\theta - 1$ agents search agent i among $\theta - 1$ agents). The variable $\nu_{ap}(\eta, \theta)$ is the union of distinct active and passive draws for i .⁶ The next proposition derives the probability distribution of $\nu_{ap}(\eta, \theta)$ and shows that it satisfied 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. Together with the decreased redundancy of i 's draws, here we have two opposite effects on i 's "passive" draws: first, as the pool grows, the set of agents searching for i grows, increasing the probability that i is found; at the same time, a larger pool makes it less likely that i is found by any given agent in the pool. This last negative effect on the probability that i is matched is responsible for the failure to rank the random variables induced by pools of different size according to First Order Stochastic Dominance.

2.2.2 Finding friends through friends

Consider a search environment in which each agent of group q is endowed with a set of "friends" belonging to the same group q . These friends are exogenously given, and can be used to search for additional friends, by randomly drawing among their own friends. Formally, we construct a network within each group by defining a set of neighbors for each agent. If average degree is assumed constant across groups independently of size, then the clustering of groups' networks is inversely related to groups' size. This implies that the probability of meeting a new friend is larger in larger groups, where the probability of redundancy, that is, of finding an agent who is already a friend, is smaller.

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), while outbreeding would mean to rely on some anonymous "market" that induces a uniform probability distribution over all agents in the population. The cost of outbreeding refers

⁶The process describe here does not describe a marriage problem, where the set of agents are actually partitioned in two subset within which agents never search one another. The same logic, however, extends to that case, with the convention that a growing pool size means that both sides of the marriage market grow in size.

here to the additional search effort required by the use of an anonymous matching technology rather than of one's own network of existing friends. This seems to well describe certain features of immigrants' social networks, where existing social ties are prominently with other immigrants, and a cost is required to expand the set of friends by opening up to local communities. It can be easily checked that the structure of equilibrium of this search game has the same qualitative features characterize in proposition 3.

2.2.3 Taste for variety

Here we think of groups as sets of agents which are close in some metric defined on relevant socio-economic characteristics, such as resource endowments (labour, land, human capital), geography, kinship, race, religion, etc... As in Dixit's (2003) model of trade, one can assume that agents are located on a circle, with distance defined as the shortest path. Agents are randomly assigned characteristics out of a given set according to some probability distribution. Probabilities are correlated with distance, in the sense that if agent i is of type A , then the probability that agent j is of type A as well is larger than the probability that agent k is of type A if and only if j is closer to i than k on the circle. Search technology is such that a cost has to be paid to search outside one's group, and agents cannot target other agents according to distance, but can decide to open up their search to the whole pool. If there are strong complementarity in production or in preferences, we can think of the random variable ν as counting the number of matches with agent of different characteristic. Then, the size monotonicity 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 finally enjoy. Denote this number by y_i for each agent i . Note that the number of enjoyed matches y_i may be related to the number of potential matches $\nu(\eta, \theta)$ in various ways. For simplicity, we assume that these two numbers coincide. Alternatively, one may assume that each meeting is itself subject to an independent probability of success, since a good matching may require that some affinity in some non modeled characteristic occurs, and this follows some given probability distribution on population. In this case the number of enjoyed matches would follow a binomial distribution. All our results carry over to this case, at the cost of much lengthier proofs.

Let now $U(y)$ denote the utility function of each agent defined on the number of enjoyed matches. We assume that U is increasing, that $U(0) = 0$ and that

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

$$U(1) \geq U(0). \quad (2)$$

The expected payoff associated with the search profile (η, θ) is then given by:

$$E_\nu U = \sum_{v_i=0}^n P(v_i) U(v_i). \quad (3)$$

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

2.4 The search game

We are now ready to define the search game in all its elements. Agents' main decision is whether to direct search towards members of his/her own group only, or to extend search to agents of different groups as well. We call the first alternative "Inbreeding", and the second "Outbreeding", denoting these choices as I and O , respectively. In doing so we are implicitly restricting the attention to the corner cases in which either $\eta_O = \eta$ or $\eta_I = \eta$. In the appendix we show that this is without loss of generality under a mild additional axiom that introduces a natural additivity requirement on the random variable ν when the pool is very large. Roughly speaking, this axiom requires that splitting up the search activity does not change the outcome of search.

A strategy profile of the search game can be therefore identified with a vector $s \equiv (s_i)_{i \in N} \in \{I, O\}^n$. For convenience, we shall restrict throughout to strategy profiles s that are group-symmetric, i.e. with $s_i = s_j$ whenever i and j belong to the same group. Thus the population behavior can be fully described by the q -tuple (s_1, s_2, \dots, s_q) that specifies a choice in the set $\{I, O\}$.⁷

Given this space of strategy profiles, the matching mechanism embodied by the function $\nu(\cdot)$, and the payoff function described in (3), in order to define the game it is enough to specify how players' strategies shape their respective search pools. Here we shall distinguish between two scenarios.

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 $\theta_l(s)$ the size of the search pool accessed in this context by group l when the (group-) strategy profile is (s_1, s_2, \dots, s_q) , we have:

$$\theta_l(s) = \begin{cases} n_l & \text{if } s_l = I \\ n & \text{if } s_l = O \end{cases} \quad (4)$$

where n_l stands for the cardinality of group l . This model can be representative of situations in which outbreeding occurs by physically going to other groups' neighbors, or learning the languages of other groups.

Such a one-sided scenario is to be contrasted with a two-sided context where the search pool of any outbreeding group consists of those groups that have themselves chosen to outbreed. This

⁷When considering a deviation from such symmetric profiles, the payoff of the deviator is still fully determined by the symmetric profile adopted by the other agents.

gives rise to an alternative function $\tilde{\theta}_l(s)$ specifying the corresponding size of the search pool, which is given by:

$$\tilde{\theta}_l(s) = \begin{cases} n_l & \text{if } s_l = I \\ \sum_{\{l': s_{l'}=O\}} n_{l'} & \text{if } s_l = O \end{cases} \quad (5)$$

This alternative scenario model may be appropriate for situations in which outbreeders move to some common location where only outbreeders meet (e.g., downtown), or learn a common third language.

Summing up, the search process can be represented as the normal form game with set of players N , set of strategies $\{I, O\}$ for each player and payoff functions $\pi_l(s) = E_{\nu(\eta, \theta_l(s))} U$.

2.5 Search Equilibrium

We now characterize the group-symmetric equilibrium of our search game. The central result is that the equilibrium behaviour of a group is fully characterized by its size, and that equilibrium strategies are of the threshold type, so that groups with size 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 prove this result in full detail for the one-sided model; for the two-sided model we focus on some additional details and rely for the proof of the one-sided model for the main arguments.

PROPOSITION 3 (Threshold Equilibrium - One-Sided Model) *Consider a one-sided search context where preferences satisfy (1)-(2). Assume that either axiom SM holds or that WM holds and the utility function U is linear. Assume also that outbreeding costs are low in the following sense:*

$$\sum_{\nu_i=0}^n (\nu_i(\eta, \infty) U(1)) > c. \quad (6)$$

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

$$\gamma_q^* = I \Leftrightarrow n_q \geq \tau. \quad (7)$$

The intuition behind this result is easy to grasp. Smaller groups are more affected by the size-monotonicity axiom, experiencing 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.

The same principle holds in the two-sided model, only that here 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 results carries over unaffected to this case.

PROPOSITION 4 (Threshold Equilibrium - Two-Sided Model) *Consider a one-sided search context where preferences satisfy (1)-(2). Assume that either axiom SM holds or that WM holds and the utility function U is linear. Assume also that outbreeding costs satisfy condition (6). Then, there exists some given (finite) $\tau \geq 2$ and \hat{n} satisfying the following property for all $n \geq \hat{n}$: if the distribution of groups' sizes is such that $\sum_{l:n_l < \tau} n_l > \alpha n$ for some $\alpha > 0$, then the strategy profile $\gamma^* = (\gamma_l^*)_{l=1}^q$ such that $\gamma_l^* = I \Leftrightarrow n_l > \tau$ for all $l = 1, \dots, q$ is an equilibrium.*

Note that in the two-sided model the strategy profile in which all groups outbreed is always a trivial equilibrium of the game. Moreover, there may be multiple equilibria in which some groups outbreed. Among these, the threshold τ^* characterizing equilibrium in proposition 3 is the largest equilibrium threshold of the two-sided search game; all other equilibria, that is, with a lower threshold, are other instances of the coordination failure underlying the trivial equilibrium in which all groups inbreed. To see this, remember that τ^* denotes the smaller group size for which inbreeding is an optimal size given that outbreeding is done in a search pool which grows unboundedly large. Moreover, by SM, search through outbreeding is most effective at τ^* . An equilibrium threshold $\tau' > \tau^*$ could only be compatible with a case in which not enough mass of small groups is present at τ^* . However, since groups larger than τ^* prefer to inbreed when outbreeding is most effective, this would be a contradiction.

3 Homophily

The equilibrium inbreeding/outbreeding choice in the search model has immediate implications for the type-composition of matches. A simple index of how "biased" actual matches are with respect to the theoretical mix implied by a uniform random assortment process is obtained by comparing the share of matches with a given type with this type's population share. More precisely, let us denote by m_i^q the number of matches that an agent of type i has with agents of type q , and by m_i the total matches of i agent. The ratio $\frac{m_i^q}{m_i}$ measures the representation of type q matches in the total matches of i . Denoting by w_q the relative population share $\frac{n_q}{n}$ of group q , we call the difference $(\frac{m_i^q}{m_i} - w_q)$ the "excess representation" of type q in the matches of agent i .

Let now $q(i)$ denote the type of agent i . When we look at the matches that agent i has with agents of the same type as i , the ratio $\frac{m_i^{q(i)}}{m_i}$ is itself an index of homophily for agent i , that we denote by H_i . We then refer to the difference $H_i - w_i > 0$ as "excess homophily" of agent i .⁸ If observed ties were generated by a uniform random assortment, then on average we would observe zero excess representation for all types and, in particular, we should observe $H_i = w_i$. An index of homophily for group q is obtained by averaging the indexes obtained for its members.

⁸The positive difference between the index H_i and the population share of group i is usually referred to as "inbreeding homophily" of group i . We do not use this terminology here in order to avoid confusion with the "inbreeding" choice of agents in our search model.

When it comes to comparing the observed homophily of different groups, however, the simple difference $H_q - w_q$ would provide a distorted picture of groups' attitudes to inbreed. In fact, groups with very large size would never experience large excess homophily, for the simple reason that the maximal potential value, measured by $1 - w$, is small to begin with. The index proposed by Colemann (1956), and recently employed in various works on the economics of homophily (see Currarini, Jackson and Pin (2009, 2010), Bramouille and Rogers (2009)) facilitates the comparison of different groups by normalizing the difference $H_q - w_q$ by the maximal value $1 - w_q$:

$$IH_q = \frac{H_q - w_q}{1 - w_q} \quad (8)$$

In the next sections we illustrate and discuss some empirical patterns of homophily (and of the Colema Index in particular) and of the distribution of inter-group ties in two diverse instances of social networks: American High School friendships and American inter-ethnic marriages. We then argue that the search model proposed in section 2 can account for such regularities.

3.1 Empirical Patterns of Homophily

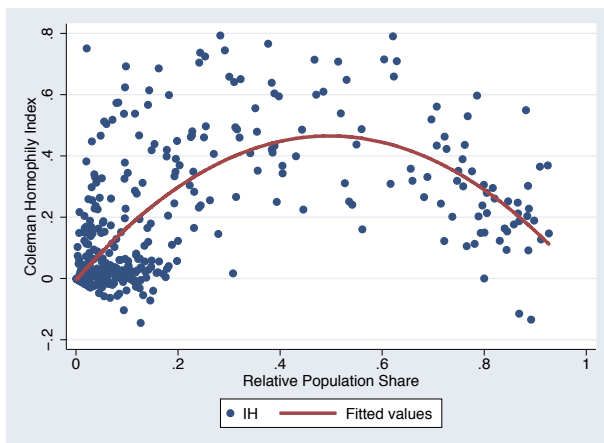
We start by reviewing some recent empirical findings by Currarini, Jackson and Pin (2009), who have studied the structure of inter-racial friendships in the Add Health sample of American High schools.⁹ They uncover a non linear, non monotonic relation between the size of racial groups and their IH index. In particular, they find that very small and very large groups display positive but very small levels, and middle sized groups display large positive values (up to about .8).

Figure 1 illustrates this relation; each dot measures the average value of the IH index for a given racial group in a given high school covered by the AddHealth sample.

We find a qualitatively very similar pattern in a different matching setting: U.S. marriages. We use the database IPUMS (Integrated Public Use Microdata Series), recording personal census data from 1850.¹⁰ An observation in our dataset identifies a triple "year, marriage market, ethnicity" (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. If a state has a city with more than 500.000 people, we split this marriage market in two: urban and rural, under the hypothesis that

⁹The 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).

¹⁰IPUMS consists of a series of compatible-format individual-level representative samples (1percent of the US population) of the American population for the years 1850-1880, 1900-2000, 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."



1: Coleman Homophily Index: High School Friendships

these two markets shows different patterns. We consider 8 ethnicities: White, Hispanic, Black, Native, Chinese, Japanese, Indian, Other Asian.

Applying the formula of the homophily indices to this case, the index H_q denotes the share of marriages involving both partners from group q on the total number of marriages that involve at least one member of group q . Population shares are here given by the share of groups' members on the total number of married people.

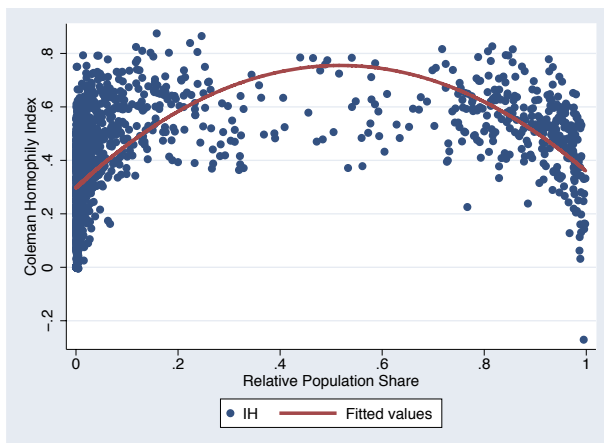
Figure 2 illustrates the behaviour of the Coleman homophily index IH as a function of the relative population shares of ethnic groups, with a pattern at all similar to the one we observed for the case of high school friendships.¹¹

The fitted line is obtained by regressing the index IH on population shares and on the square of population shares.¹² We obtain similar patterns for subsamples referring to single years and to single ethnic groups, that we omit for expositional convenience.

We show in the next section that our search model generates such a non monotonic trend of the Coleman Index once small search frictions are built in the search technology. This result adds up to previous explanations of such trend, given in Currarini, Jackson and Pin (2009) and in Rogers and Bramouille (2009), and based on different models of search. Before turning to a formal proof of this result, we document some additional empirical evidence that supports the qualitative features of the threshold equilibrium, specific to our model of search and proved in propositions 3 and 4. This additional evidence is obtained by decomposing by race the total number of inter-ethnic ties that a racial group has. Figure 3 makes use of this decomposition to illustrate the distribution of inter-

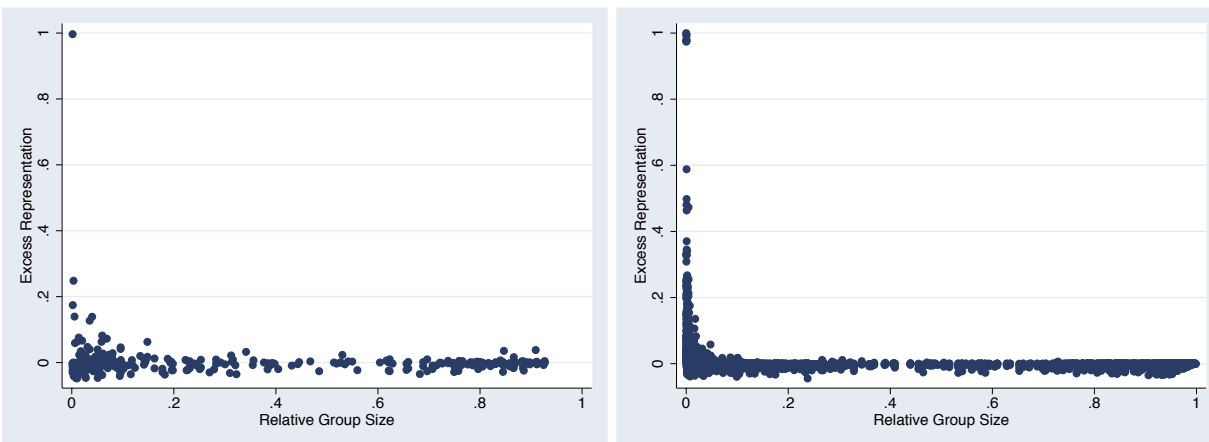
¹¹The main difference between figures 1 and 2 is that the intercept of the regressed line is significantly different from zero in the case of marriages. This point was brought up in a recent paper by Franz, Marsili and Pin (2008), where the value of the intercept is used to assess the presence of biases in the meeting process.

¹²Details of the regression are as follows: $H_q = 0.262 + 0.768\sqrt{w_q}$
(.003) (.003)



2: Coleman Homophily Index: Marriages

ethnic ties that are formed with members of "small" groups, both in the high school friendships and in the marriages datasets.¹³ On the vertical axis we measure the "excess representation" of groups of 5 per cent or less of the total population in the ties formed by groups whose population share is measured on the horizontal axis.



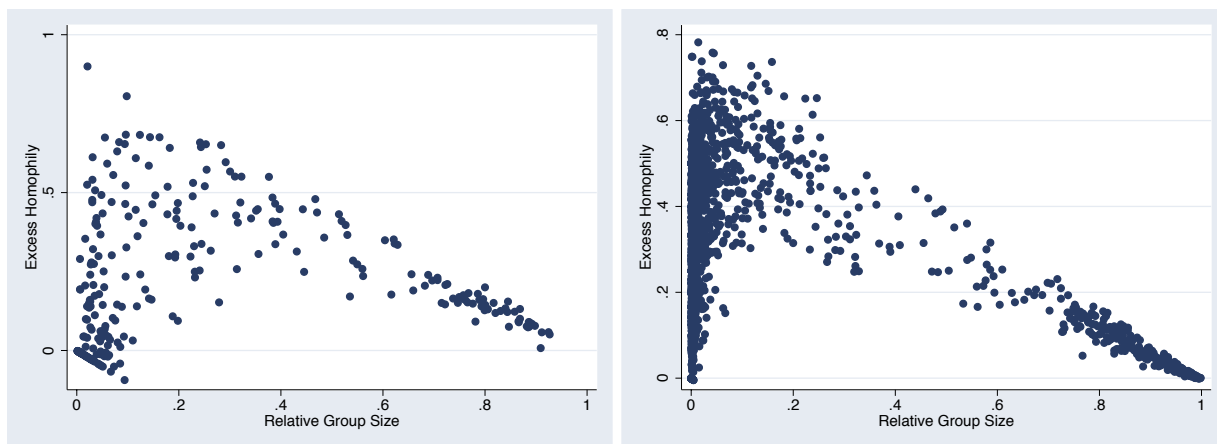
3: Inter-ethnic Ties with Small groups: High School Friendships (left) and Marriages (right).

We observe that in both friendships and marriages, small groups tend to be over-represented in the ties formed by other very small groups (small values on the horizontal axis), while they either are under-represented or reflect their population share for medium and large groups. In particular, and most clearly in the marriage dataset, there seems to be some very small critical size of groups after which the over-representation of small groups disappears. These pictures are compatible with

¹³The friendship data refer here only to reciprocated nominations.

an inbreeding/outbreeding behavior of groups that follows the threshold equilibrium property of our model. In particular, if the tie generative process is described by the "two-sided" variant of our model (see proposition 4), then outbreeding groups (whose size is below the equilibrium threshold) end up uniformly searching over the pool of all outbreeders, finding all outbreeding groups (including themselves) at a rate that exceeds the relative population shares. This is reflected by the positive values (measured on the vertical axis) that correspond to small groups on the horizontal axis in figure 4.

Note also that our two-sided search model predicts that the excess homophily of groups should jump upwards at the equilibrium threshold, and then decrease linearly after that, approaching zero for very large groups. In Figure 4 we see that this is approximately the trend followed in both marriages and high school friendships.



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

3.2 Coleman homophily index by group size: equilibrium predictions

We now characterize, in a series of propositions, the equilibrium relation between group size and the Coleman homophily index.

We start by expressing the variables involved in the formulation of the Coleman Index in terms of the parameters that characterize our search model. For a given pool size θ , the expected total number of matches is given by:

$$m(\eta, \theta) = \mathbb{E}[\nu(\eta, \theta)]. \quad (9)$$

The expected number of same group matches for an agent of an inbreeding group is exactly $m(\eta, \theta)$; outbreeding groups have instead proportions of same and different group matches that follow population shares. The expected number of same group matches for an agent of an outbreeding group of relative size is therefore $w_q m(\eta, \theta)$.

Applying the formula for the Coleman Index to a group of relative size w , we obtain the following values as a function of the outbreeding/inbreeding decision:

$$\begin{aligned}
IH(I) &= \frac{1 - w_q}{1 - w_q} = 1; \\
IH(O) &= \frac{\frac{w_q m(\eta, \theta)}{m(\eta, \theta)} - w_q}{1 - w_q} = 0.
\end{aligned}$$

Given the threshold structure of the search equilibrium, we obtain a very simple pattern for the IH index with respect to group size: zero up to the threshold level τ , and 1 afterwards. In the rest of this section we show how by building small frictions into the search model studied above we can generate a richer pattern of the IH index which is consistent with the empirical evidence discussed in section 3.1. We assume that agents are exposed to meeting opportunities that are beyond their inbreeding/outbreeding control, but that are a pure result of the randomness of social interaction. In particular, we do this by defining by r_I and r_O additional search efforts that each agent exerts within his own group and in other groups, respectively.¹⁴

We start by showing that very small groups, and in particular outbreeding groups, are characterized by a very small Coleman Index as population size grows large. In what follows, we denote by C_q the Coleman Index of group q .

PROPOSITION 5 *Consider any group q such that $n_q < \tau$, where τ is as specified in Proposition 3. Then, there exists some \hat{n} such that if $n \geq \hat{n}$, then C_q is bounded above by the term $\frac{m(r_I, n_I)}{m(\eta + r_O, \infty)}$.*

The next proposition shows that outbreeding groups have Coleman Indices that are ordered according to size. Since outbreeding groups are small, this result matches the empirical observation that the Coleman Index is sharply increasing for low values of group size.

PROPOSITION 6 *There exist some \hat{n} such that if $n \geq \hat{n}$, any two outbreeding groups q and q' with $n_q < n_{q'}$ satisfy $C_q \leq C_{q'}$.*

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 high levels of the Coleman Index for low levels of search frictions.

PROPOSITION 7 *Given any $\varepsilon > 0$, there exist some $\delta_1 > 0$, $\delta_2 > 0$ and \hat{n} such that if $n \geq \hat{n}$ and $\frac{m(r_O, \infty)}{m(\eta + r_I, \infty)} \leq \delta_1$ then any group q with relative size $\delta_1 > \frac{n_q}{n} > \delta_2$ has $C_q \geq 1 - \varepsilon$.*

¹⁴If we think of search effort as time, then r_I and r_O can be thought as units of time during which social encounters can happen independently of the intention of an agent to devote that time to search. Note that in this case, matches with other groups come free of the outbreeding cost c .

Finally, the next two results complete the previous analysis by specifying how the homophily index changes with group size among relatively large groups that inbreed. First, Proposition 8 shows that, among groups that inbreed and have a nonnegligible relative size, their IH index decreases as they grow larger. Second, Proposition 9 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. These results qualitatively match the empirical evidence of a decreasing IH index in the range of very large groups, with negative values for groups of size close to the whole population.

PROPOSITION 8 *Consider any two groups, q and q' , such that their relative sizes are bounded away from 1 and such that $\frac{n_{q'}}{n} - \delta_1 \geq \frac{n_q}{n} > \delta_2$ for some $\delta_1, \delta_2 > 0$. Then, there exists some \hat{n} such that if $n \geq \hat{n}$, $C_q < C_{q'}$.*

PROPOSITION 9 *There exist some $\delta_2 > 0$ and \hat{n} such that if $n \geq \hat{n}$, then any group q with relative size $\frac{n_q}{n} \geq 1 - \delta_2$ has $C_q < 0$.*

Summing up, the various propositions proved above identify four classes of group size, whose equilibrium behaviour (and the associated Coleman index) is consistent with the empirical evidence discussed in section 3.1.

1) **Small groups.** These are groups whose relative size gets arbitrarily close to zero as n grows. These groups may either outbreed or inbreed. Among these, the smallest outbreed, and their Coleman index is increasing in group size.

2) **Intermediate groups.** These are groups whose size is bounded away from both zero and 1 as n grows. These groups always inbreed, and their Coleman index is arbitrarily close to 1, decreasing with group size. The closer to 1 we want their Coleman index, the closer we must set the size r_O of the search friction.

3) **Large groups.** These are groups whose relative size is bounded away from both zero and one as n grows. Their Coleman index is decreasing in group size.

4) **Very large groups.** These groups cover an arbitrarily large proportion of the population, and have IH indexes that become negative in the limit.

4 Concluding Remarks

In a model where agents are organized in groups that are homogeneous with respect to some relevant characteristics, we have studied the incentives of agents to mix with other groups, and how these incentives vary with group size. In doing this we are motivated by consistent empirical evidence across applications that group size matter in determining the homophilous behavior of groups, by affecting the attitude of their members towards other groups. The key ingredient of our approach is the assumption that search for social contacts is more effective when exerted in larger pools.

This assumption is meant to capture forces that may be at work in certain contexts of network formation, and that may generate from redundancy in search and from taste for variety, as we have suggested in Section 2.2. The immediate consequence is that small groups face higher incentives to open up and mix with other groups, since in-group search turns out to be little effective.

While all our analysis was centered on homophily, there is one additional empirical regularity that Currarini, Jackson and Pin (2009) have shown to characterize high school friendships, and to be related to homophilous preferences: larger groups make more friends on average. While our model is not rich enough to generate the feedback mechanism between choice and opportunities that drives this pattern in the CJP paper, this empirical pattern would occur under some conditions in the equilibrium of the "two-sided" version of our model. In particular, if the mass of outbreeders is not large enough to generate the incentives to bear the cost of outbreeding, all groups inbreeding in equilibrium. In this case, members of larger groups search more effectiveness (simply because their search pool is larger), and end up finding more matches on average.

References

- [1] Allport, W. G. (1954): *The Nature of Prejudice*. Cambridge, MA: Addison-Wesley.
- [2] Blau, P. M. (1977), *Inequality and Heterogeneity: A Primitive Theory of Social Structure*. New York: Free Press.
- [3] Bramouillè, P. and B. Rogers (2009) "Diversity and Popularity in Social Networks", mimeo.
- [4] Coleman, J. (1958): Relational Analysis: The Study of Social Organizations With Survey Methods, *Human Organization* 17, 2836.
- [5] Currarini, S., M.O. Jackson, and P. Pin (2009) "An Economic Model of Friendship: Homophily, Minorities and Segregation," *Econometrica* 77, No. 4, 1003–1045.
- [6] Currarini, S., M.O. Jackson, and P. Pin (2010) "Identifying the Roles of Choice and Chance in Network Formation: Racial Biases in High School Friendships", *Proceedings of the National Academy of Science* 107, 4857-4861.
- [7] Dixit, A. (2003), "Trade Expansion and Contract Enforcement", *Journal of Political Economy* 111(6), 1293- 1317.
- [8] Franz, M. Marsili and P. Pin (2008), "Observed choices and underlying opportunities", mimeo.
- [9] Giles, M. W. (1978), "White Enrollment Stability and School Desegregation: A Two- Level Analysis" *American Sociological Review* 43, 2448-64.

- [10] McPherson, M., L. Smith-Lovin and J. M. Cook (2001): Birds of a Feather: Homophily in Social Networks, *Annual Review Sociology* 27, 415-444.
- [11] Moody, J., "Race, School Integration, and Friendship Segregation in America" *The American Journal of Sociology*, 107(3), 679-716.
- [12] Stajic, W. (1990), "The Collector's Problem with Group Drawings ", *Advances in Applied Probability*, 22(4), 866-882.

Appendix

ADD (Additivity in Search) For all (η_O, η_I) such that $\eta_O + \eta_I = \eta$, let $\nu(\eta_O, \eta_I, \infty) \equiv \nu(\eta_O, \infty) + \nu(\eta_I, \infty)$ be the random variable associated with the sum of the outcomes of two search processes with intensities η_O and η_I on an infinite pool. Then $\nu(\eta_O, \eta_I, \infty) = \nu(\eta, \infty)$.

LEMMA 1 *If the random variable ν satisfies either axioms SM and ADD, and if the utility function U is monotone increasing, then, for large enough n the optimal choice of each agent either involves $\eta(I) = \eta$ or $\eta(O) = \eta$.*

Proof of Lemma 1 By SM, $\nu(x, \infty)$ FSD $\nu(x, \theta)$ for all x and θ . So, since the outbreeding pool is always weakly larger than the inbreeding pool, and for large enough n it is infinite, then we have that $\nu(\eta_I, \theta_I) + \nu(\eta_O, \infty)$ is dominated by $\nu(\eta_I, \infty) + \nu(\eta_O, \infty)$, which is equal to $\nu(\eta, \infty)$. Since the cost is fixed, this, together with the fact that U is increasing, implies that an agent prefers to outbreed for η units of time than for η_O units of time, for all $\eta_O \leq \eta$.

Proof of Proposition 1

Let us denote by $p(d, \eta, \theta)$ the probability of d distinct draws from η draws with replacement out of a pool of size θ . 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 $p(d, \eta, \theta + 1)$:

$$\frac{p(d, \eta, \theta)}{p(d, \eta, \theta + 1)} = \frac{\binom{\theta}{d} \sum_{j=0}^d (-1)^j \binom{d}{j} \left(\frac{d-j}{\theta}\right)^\eta}{\binom{\theta+1}{d} \sum_{j=0}^d (-1)^j \binom{d}{j} \left(\frac{d-j}{\theta+1}\right)^\eta} \quad (10)$$

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 θ 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 η 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:

$$(\theta - l) \cdot p(\eta, \theta) \tag{11}$$

where

$$p(\eta, \theta) = \left(1 - \left(\frac{\theta - 1}{\theta}\right)^\eta\right). \tag{12}$$

is the probability that an agent in a pool of size θ is found by means of η 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 (11) over all possible values of l using the binomial distribution. We obtain the following:

$$E(\nu_{ap}(\eta, \theta)) = \sum_{l=0}^{\theta} \binom{\theta}{l} p(\eta, \theta)^l (1 - p(\eta, \theta))^{\theta-l} \cdot l + \sum_{l=0}^{\theta} \binom{\theta}{l} p(\eta, \theta)^l (1 - p(\eta, \theta))^{\theta-l} \cdot (\theta - l) \cdot p(\eta, \theta). \tag{13}$$

We can rewrite the second sum in (13) as follows by extracting the term $\theta \cdot p(\eta, \theta)$ which does not depend on l :

$$\theta \cdot p(\eta, \theta) \sum_{l=0}^{\theta} \binom{\theta}{l} p(\eta, \theta)^l (1 - p(\eta, \theta))^{\theta-l} - p(\eta, \theta) \sum_{l=0}^{\theta} \binom{\theta}{l} p(\eta, \theta)^l (1 - p(\eta, \theta))^{\theta-l} l. \tag{14}$$

Note that the second term of (14) 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 θ is given by:

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

and the sign of $\frac{\partial f(\theta, \eta)}{\partial \theta}$ 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 2η times the increase of the log function from $\theta - 1$ to θ 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 in a group-symmetric profile s is independent of the choice of groups other than l . The expected payoff $\pi_l(s)$ for an individual i of an outbreeding group l is given, for large n , by the expression:

$$\pi_O(n_l) = \sum_{\nu_i=0}^n P(\nu_i(\eta, \infty)) U(\nu_i) - c - slo(n), \quad (15)$$

where $slo(n)$ is a nonnegative infinitesimal in n . Since $U(0) = 0$ and $U(y) \geq U(1)$ for all $y \geq 1$, we can write:

$$\pi_O(n_l) \geq \sum_{\nu_i=0}^n (\nu_i(\eta, \infty) U(1) - c - slo(n)). \quad (16)$$

Consider now the payoff $\pi_I(n_l)$ that agent i gets if she inbreeds. For $n_l = 2$, this payoff is simply

$$\pi_I(n_l) \equiv p U(1) \quad (17)$$

so that, from the assumptions of the Proposition, we can guarantee that $\pi_O(n) > \pi_I(n_l)$ if n is large enough, in which case it is optimal for any such agent to outbreed.

Using now SM and the fact that $U(y + 1) - U(y) \geq 0$ for all y , with the inequality being strict for $y = 0$, it follows that

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

for all $n_l \geq 2$. The same result holds under the weaker WM axiom if $U(y_i)$ is assumed linear.

Let now τ be the lowest integer n_l such that

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

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 $\nu_i(\alpha n)$ approaches the limit distribution $\nu_i(\eta, \infty)$. This implies that (15) holds with the term $o(n)$ vanishing. The same steps as in proposition 4 can be therefore used to prove that the same threshold found in proposition 4 satisfies the conditions for the present result.

Proof of Proposition 5 First note that, since τ is independent of n , then $\frac{n_q}{n} \searrow 0$ as $n \nearrow \infty$. Thus, for n large enough C_q can be approximated by $\frac{s_q}{s_q + d_q}$. Let us now consider the expected number of same type matches s_q for group q . Consider the probability measure associated with the random variable $\nu(\eta, \theta)$ and assigning to each set of agents S the probability that $\nu(\eta, \theta) \cap S \neq \emptyset$. Provided that this measure is absolutely continuous with respect to the fractional measure defined on the sigma algebra of N , then if group q outbreeds and $\frac{n_q}{n} \searrow 0$ for large enough n , we can approximate the expected value of s_q by $m(r_I, n_q)$, and d_q by $m(\eta + r_O, \infty)$. Then we obtain:

$$C_q = \frac{s_q}{s_q + d_q} = \frac{m(r_I, n_q)}{m(\eta + r_O, \infty)}.$$

Proof of Proposition 6 For simplicity, consider two outbreeding groups q and q' whose cardinality differ in just one individual, i.e. $n_{q'} = n_q + 1$, and let $\Delta \equiv C_{q'} - C_q$ denote the change in the Coleman index. Since passing from q to q' 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. By the SM axiom (or its weaker version WM), $m(r_I, n_{q'}) - m(r_I, n_q) > 0$. Let us write the change Ξ in the numerator of the Coleman Index when passing from n_q to $n_q + 1$ as follows:

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

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

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

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

is strictly positive, and is bounded away from zero uniformly for all n_q and $n_{q'}$ smaller than τ . It is clear that for n large enough, Ξ is strictly positive.

Proof of Proposition 7 A preliminary observation is that, if n is large enough, then since $\frac{n_q}{n}$ is bounded away from zero by δ_2 it must be that $n_q \geq \tau$ (where τ is as in Proposition 3) and therefore group q inbreeds. Moreover, for large n its size can be so large that its ratio $\frac{s_q}{s_q+d_q}$ 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_q to be approximated as follows:

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

An appropriate choice of δ_1 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 8 Consider two groups, q and q' , whose relative sizes are bounded below by some positive number δ_2 . As n becomes large, both groups must exceed the threshold τ 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 $\frac{n_{q'}}{n} - \frac{n_q}{n} \geq \delta_1$ for some positive δ_1), the desired conclusion reads:

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

or

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

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

Proof of Proposition 9 Given η , 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 $1 - \frac{\delta_2}{2} \geq \frac{n_q}{n} \geq 1 - \delta_2$ then C_q – whose sign is the sign of the term $\frac{m(\eta+r_I,\infty)}{m(\eta+r_I,\infty)+m(r_O,n-n_q)} - \frac{n_q}{n}$ for large n – is negative. To see this, note that $m(r_O,\infty) > m(r_O,n-n_q)$ for all n_q , and that, if δ_2 is in the assumed range, then:

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