Diversity and Popularity in Social Networks

Yann Bramoullé and Brian W. Rogers*

August 31, 2010

Abstract: We present a new model to understand the nature and structure of homophily in social networks. We introduce heterogeneity to the study of stochastic network formation and show that meeting friends of friends tends to reduce social segregation. This allows us to derive sharp theoretical predictions on homophily patterns. Our main result is that individuals with more connections tend to have more diverse networks. Our theoretical predictions are well supported empirically when looking at friendships between boys and girls in U.S. high schools. We derive some welfare implications and show that our analysis can help identify the source of homophily.
JEL Codes: A14, D85, I21.

*Bramoullé: Department of Economics, CIRPÉE and GREEN, Université Laval.
Rogers: MEDS, Kellogg School of Management, Northwestern University.
Vincent Boucher provided excellent research assistance. We thank Habiba Djebbari, Andrea Galeotti, Sanjeev Goyal, Matthew Jackson, James Moody, Betsy Sinclair, Bruno Strulovici, and Adrien Vigier, as well as seminar participants at the First Transatlantic Theory Workshop (Paris), the DIME conference on the formation and the evolution of social and economic networks (Paris), the University of Essex, Cambridge University, Oxford University, EUI, University of Venice, Bocconi University, and the Paris School of Economics for particularly helpful comments and suggestions. This research uses data from Add Health, a program project designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris, and funded by a grant P01-HD31921 from the National Institute of Child Health and Human Development. Persons interested in obtaining data files from Add Health should contact Add Health, Carolina Population Center, 123 W. Franklin Street, Chapel Hill, NC 27516-2524 (addhealth@unc.edu).

I Introduction

Economists are becoming increasingly aware of the importance and ubiquity of social networks. The structure and properties of social networks capture important information on individual preferences and opportunities. In turn, these networks determine social and economic outcomes as diverse as information diffusion, disease transmission, informal insurance, access to jobs, and risky behavior.[1]

Fine details of the network's architecture, such as the prevalence, characteristics and positions of social hubs, may have major impacts on these outcomes. The extent to which similar individuals tend to be connected, or *homophily*, is particularly informative on the extent of social segregation. At this stage, homophily is the major issue in need of deeper analysis. Despite its empirical prevalence, we currently have a poor understanding of its causes, consequences and interplay with other network features.

We present a new model to understand the nature and structure of homophily in social networks. A main feature of our approach is that it generates *empirically realistic* networks. To achieve this we introduce heterogeneity to the study of growing networks. Our framework gives rise to networks which satisfy key statistical properties of real social networks while allowing the analysis to focus on understanding homophily.[2]

We develop our analysis in four directions. First, we provide a thorough theoretical study of the model's properties with respect to homophily and its interactions with other network properties. Most importantly, we show that the model yields a negative homophily-degree relationship: Individuals who are more connected tend to have more diverse neighborhoods.

[1]The literatures on the impact of social network on these various outcomes is large. For specific examples, see Conley and Udry (2001, 2010), Laumann and Youm (1999), Fafchamps and Lund (2003), Bramoullé and Kranton (2007), Ambrus, Mobius and Szeidl (2010), Beaman (2010), Montgomery (1991), Calvo and Jackson (2004), Christakis and Fowler (2008).

[2]Most importantly, networks generated by our model satisfy the following five properties, all of which are empirically documented (see Jackson & Rogers (2007)). The distribution of degree is scale-free in the upper tail, and thinner in the lower tail. There is high clustering of links and many short paths between individuals, so the network will be a "small world." The network is assortative in the sense that high-degree individuals have high-degree neighbors. Finally, high-degree individuals have lower clustering among their neighbors than low-degree individuals.

Second, we confront our theoretical predictions with data. Using the AddHealth data, we find that in U.S. high schools boys and girls who receive more friendship nominations indeed have a much more gender-diverse network of friends. Third, we use our model to provide some simple welfare analysis. We find conditions under which an optimal level of homophily emerges, and study how this level varies between groups and with parameters of the model. Fourth, we show how our framework can help identify the source of homophily and especially how it can help separate institutionally-driven from preference-driven biases.

Briefly, the framework we study works as follows. As in Jackson & Rogers (2007), we model a growing population of individuals who develop links to each other over time. Individuals are born sequentially. Upon entry, an individual meets others in two ways: at random and through network-based search. The network-based search allows individuals to meet neighbors of their randomly met partners, or "friends of friends". In either case, conditional on a meeting taking place, a connection is formed with some probability. Our point of departure is that individuals exogenously belong to one of two groups (e.g., boys and girls) and that group identity affects the original random meetings. We study how this bias propagates throughout the system as the population grows large.

We uncover a key new mechanism which determines the structure of homophily in our setup. Namely, that *network-based search tends to reduce homophily.* This mechanism is not obvious a priori, given the presumption that social segregation may be aggravated by network effects. Yet it has an intuitive explanation. Even if meetings across groups are relatively rare, a new individual still meets *some* individuals from the other group. Through them, he then has access to many additional individuals from the other group, and so diversity among friends of friends is greater than among friends. We believe that this captures an important mechanism at work in real networks, and one that has not been previously identified or studied.

In the theoretical part of our analysis, we clarify the consequences of this central finding. We show, first, that the aggregate level of homophily in the network decreases as the proportion of search-based meetings increases. This result ties the consequence of observed homophily to the

details of the how links are formed at the individual level.

Second, and more importantly, we examine how the popularity of an individual is related to her local homophily, i.e., the proportion of her neighbors belonging to the same group. We find that *local homophily decreases with an individual's degree*. The reason is that individuals with high degree are relatively more likely to be found through the network. So the diversity-enhancing effect of network search leads social hubs to be matched with a more diverse population of potential partners. Further, we find that this decrease takes place at a decreasing rate, that it is more pronounced for individuals in smaller groups, and that the homophilic bias completely disappears in the limit as degree becomes large. This analysis provides the first comprehensive results relating homophily to degree in social networks. We feel that such results, which help to better understand the characteristics of the most highly connected individuals, are particularly important in light of the fact that social hubs have a disproportionate effect on network processes such as communication and diffusion.

Our theoretical analysis yields a number of additional non-trivial testable implications. Specifically, we look at the relationship between the relative homophily within a particular group and the relative size of that group. As in Currarini, Jackson & Pin (2008), we find that this relationship exhibits an inverted-U shape. It is remarkable that two models built on very different premises lead to the same conclusion, and leads us to believe this conclusion is particularly robust. Finally, we characterize the degree distributions of the social networks within and across groups, and derive a number of comparisons across them using first-order stochastic dominance.

We then bring data to bear on these predictions. We look at gender biases in friendship networks among high school students in the U.S. provided in the AddHealth data. We test the hypotheses that local homophily is decreasing and convex in degree, and that the relationship is more dramatic in smaller groups. Regressions demonstrate highly significant relationships between our variables of interest, all with the predicted signs. This exercise constitutes the first empirical analysis of the relationship between homophily and degree, and provides strong

support for the model's predictions. More broadly, our framework could guide future empirical research on gender bias in other settings, including gender composition and social interactions at school, scientific collaboration patterns and professional interactions, see e.g. Boschini & Sjögren (2007).

We next derive some welfare implications of our analysis. We study how the preferences of individuals over their friends induce preferences on the network formation process. Under a natural specification of preferences, we find that the minority group always prefers a more biased linking process than the majority group. We also uncover the possibility average utility varies non-monotonically with the ratio of random to search-based meetings, so that there may be an interior optimal level of network-based search. Last, we tie the negative homophily-degree relationship to the rate at which utility increases with degree. Again, there is the possibility of non-monotonicities.

Finally, our analysis makes progress towards the important issue of understanding the origins of homophily. In particular, we provide a clear picture of the structure of homophily induced by bias in random meetings. We interpret the bias that we study as reflecting features of the social environment that result in meeting similar individuals disproportionately to their representation in society. Other biases may also be significant. For example, under a bias in type-based preferences, individuals may be more likely to maintain relationships with similar individuals conditional on meeting them. Different sources of bias imply different structural features in the network and, in principle, they can be identified on that basis. In our context, we show that *a pure bias in preferences leads to a flat homophily-degree relationship*, and hence cannot explain the empirical findings. Camargo, Stinebrickner and Stinebrickner (2010) provide further evidence that random meeting bias is more important than preference bias in explaining racial homophily in college friendships, by finding that randomly assigned roommates of different races are as likely to become friends as randomly assigned roommates of the same race. Both our work and this empirical evidence complements recent work by Currarini, Jackson & Pin (2010), who obtain empirical estimates of the magnitude of two similar biases in the context of their very different model, see also Mayer and Puller (2008). Their identification strategy exploits

variations in average degree and average homophily between the various groups. We show here that finer structural features of homophily also have identifying power.

While our model is stochastic in nature, it can be given simple microfoundations. We think that economists should not shy away from these models because of such simplicity. Stochastic models currently provide, by far, the best available methodology with which to study social networks empirically.[3] Their explanatory power comes largely from the way natural social processes are built into them. As in Jackson & Rogers (2007), our model formally captures the notion that people often meet one another through common acquaintances ("friends of friends"). There is strong evidence that this mechanism plays an important role in friendships. For example, Camargo, Stinebrickner and Stinebrickner (2010) show that white college students who are randomly assigned a black roommate become friends with more black students than white students who are randomly assigned a white roommate. While a fully rational modeling of such mechanisms may be feasible and desirable, it is not needed to make use of these models as effective "reduced-form"representations of behavior. In addition, stochastic models are analytically tractable, and hence can be easily incorporated into further studies of network effects, as in Golub & Jackson (2008) who study communication across networks, or welfare impacts, as studied here.

Given its nature, this study advances the analysis of stochastic models of network formation. Earlier work has made great progress in explaining structural network features such as small diameter, high clustering and fat tails in degree distributions, see Barabasi & Albert (1999), Chung & Lu (2002a,b), Jackson & Rogers (2007), Newman (2003, 2004), and Watts & Strogatz (1998). However, most of these studies assume homogeneous agents and neglect homophily. With respect to this literature, we develop and study one of the first stochastic model of network formation incorporating individual heterogeneity. We investigate the interplay of homophily with the other network features.

[3]The few existing econometric studies of network formation focus on bilateral links and neglect structural features, see Bramoullé and Fortin (2010) for an overview. Strategic models of network formation raise formidable theoretical difficulties. They still await a proper empirical implementation, yet alone one that could explain empirical network features.

Our analysis also contributes to a nascent economic literature interested specifically in homophily.[4] As mentioned, Currarini, Jackson & Pin (2008, 2010) study a matching process of friendship formation. They document several empirical patterns of homophily and explain them through a combination of biases with respect to choice (preferences) and chance (opportunities presented by the matching process). By design, their model does not allow degree to vary across individuals. This makes their entire analysis very different from ours. Interestingly, however, a number of the predictions from their model are also supported by our work. Jackson (2008a) incorporates homophily into the random graph model of Chung & Lu (2002a,b).[5] Again by design, homophily does not vary with degree in this approach. Moreover, degree distributions constitute an outcome of our model while they are an input of Jackson's (2008) analysis. Thus, our approach and these two papers study homophily patterns in networks from three distinct and rather complementary points of view. In particular, we provide the first study of the relationship between homophily and an individual's degree.

The paper proceeds as follows. The next section describes the formal model. Section III presents results on average homophily. Section IV describes how local homophily varies with degree, which highlights the importance of studying a model with non-trivial degree distributions. Section V derives degree distributions for the network under the mean-field approximation, and presents a number of relationships across degree distributions. Section VI presents the results of the empirical analysis of homophily in networks, and relates the findings to the predictions of our model. Section VII studies welfare. Section VIII introduces different kinds of biases to the model, and discusses how they can be empirically identified. The final section offers concluding remarks. An appendix contains supplemental formal results.

[4]The study of homophily dates back to Lazarsfeld & Merton (1954); McPherson, Smith-Lovin & Cook (2001) document observations of homophilic biases in dozens of studies.

[5]Golub & Jackson (2008) use this extension to study how homophily affects communication dynamics in networks, demonstrating explicitly one way in which homophilic structure impacts outcomes.

II A model of community structure in social networks

We introduce the notion of community structure to a version of the model from Jackson & Rogers (2007), which we now describe.

Nodes enter the world one at a time and are indexed by their date of birth $t = 1, 2, ...$ Entering nodes meet some of the existing nodes via two processes: at random and through network-based search. Some of these meetings result in a relationship, modeled as a directed link from the new node to the older node. In this case we refer to the older node as an "out-neighbor" of the entering node, and conversely to the entering node as an "in-neighbor" of the older node.

Specifically, each entering node first meets m_r existing nodes through the random meeting process, where the meetings are drawn uniformly and independently at random from the set of existing nodes. We sometimes refer to nodes so met as *parents*. It then meets an additional m_s nodes chosen uniformly and independently at random from the set of nodes consisting of the union of parents' out-neighbors. Finally, it independently forms a directed out-link with probability p with each of these $m_r + m_s$ nodes that it has met. The expected number of links thus formed is $m = p(m_s + m_r)$; $r = m_r/m_s$ represents the ratio of the number of connections formed through the random process versus through the search process.

Our key assumption is that individuals are endowed with types, which we refer to as group membership, and that group identity (potentially) impacts the meeting process. Specifically, nodes belong to one of two groups g^1 and g^2. At birth they are independently assigned to g^1 with probability q and to g^2 with probability $1 - q$. Thus at time t, the expected number of nodes in g^1 is qt and in g^2 is $(1 - q)t$.

We suppose that there are two locations L^1 and L^2. Each node that enters goes to one location, and meets m_r nodes uniformly at random among all individuals present at this location. All biases in the meeting process are captured by the parameter $\gamma \in [1/2, 1]$, which represents the probability that a g^i node goes to location L^i, $i = 1, 2$.[6] Thus, it is simply the resulting

[6]Actually, the analysis below assumes away some implicit correlations in the meetings process described here.

composition of types in the two locations that permits any group-dependent biases in the model. Once at a particular location, random meetings ignore types. In addition, both the search-based meetings and the probabilities of forming a link conditional on a meeting taking place ignore types and are exactly as described above.

At any time t, the expected number of g^1 nodes at L^1 is $q\gamma t$, while the expected number of g^2 nodes in L^1 is $(1-q)(1-\gamma)t$. Thus, the proportion of g^1 nodes in L^1 is $\frac{q\gamma}{q\gamma+(1-q)(1-\gamma)}$ while the proportion of g^1 nodes in L^2 is $\frac{q(1-\gamma)}{q(1-\gamma)+(1-q)\gamma}$. Defining b_i as the proportion of a g^i node's random meetings that are within its own group, $i = 1, 2$, this yields:

$$b_1 = \gamma\frac{q\gamma}{q\gamma + (1-q)(1-\gamma)} + (1-\gamma)\frac{q(1-\gamma)}{q(1-\gamma) + (1-q)\gamma} \tag{II.1}$$

and b_2 is obtained by symmetry exchanging q and $1-q$.

Once the population sizes, q, have been fixed, a single parameter, γ, determines how the random meeting biases, b_1 and b_2, vary. However, we prove many of our results in a more general setting in which b_1 and b_2 are taken as primitives of the model rather than implications of location-based biases and can vary freely with respect to each other. In these cases, the process can be viewed as follows. Independently for each random meeting of an entering node of type g^i, a coin is first flipped to determine the group of the individual to be met, with probability b_i of choosing another individual from g^i. Then, conditional on the group, an individual is selected uniformly at random from the appropriate type.

Whether or not the meeting process is location-driven, we can think of any resulting biases as being described by b_1 and b_2. Moreover we must interpret b_1 and b_2 relative to q, which measures the relative proportions of the two groups. In particular, we say that there is *no homophilic bias* if $b_1 = q$ and $b_2 = 1-q$. In that case, the proportions of random meetings within and between the groups simply reflects the relative sizes of the groups in the population. When biases are driven by locations, this corresponds to $\gamma = \frac{1}{2}$. In contrast, we say that there is *homophilic bias*

To account for this, one can instead assume that each new node in g^i spends a proportion γ of his time in L^i, and the probability of meeting any existing node is proportional to the time spent with it in the same location.

if $b_1 > q$ and $b_2 > 1 - q$ and random meetings are relatively biased towards own group. This occurs whenever $\gamma > \frac{1}{2}$. At the extreme when $\gamma = 1$, so that $b_1 = b_2 = 1$, individuals meet others only from their own group.[7]

III Homophily in relationships

First, we want to understand how homophilic bias in random meetings relates to the homophily observed in the resulting network. In particular, for a given specification, does increasing the proportion of search-based meetings amplify or mitigate the homophilic bias?

Formally, define *group i network homophily,* β_i, as the expected proportion of the links formed by a new node in g^i that are with other nodes in g^i. A priori β_i depends on t, but we can show that β_i quickly reaches a steady-state, in which case β_i measures the proportion of the links formed by all nodes in g^i that point to nodes also in g^i. Given values of b_1 and b_2, β_1 and β_2 must solve the following system of equations at the steady-state:

$$\begin{aligned} \beta_1 m &= p\left[b_1 m_r + (\beta_1 b_1 + (1-\beta_2)(1-b_1))m_s\right] \\ \beta_2 m &= p\left[b_2 m_r + (\beta_2 b_2 + (1-\beta_1)(1-b_2))m_s\right]. \end{aligned}$$

To see why, consider, for instance, the first equation. The left hand-side is the expected number of links that a new node in g^1 forms with existing nodes in g^1. On the right hand side, this number is expressed as the sum of the expected number of links in g^1 formed at random (pb_1m_r) and through search. Among all the parents' out-neighbors, what proportion of them are in g^1? A fraction b_1 of parents are in g^1, and the proportion of their out-neighbors in g^1 is (by definition) β_1, while the remaining proportion $1 - b_1$ of parents in g^2 have a proportion $1 - \beta_2$ of their out-neighbors in g^1. Since the new node has m_s search-based meetings, we get the second term in the equation. Rewriting these equations as functions of the ratio of random to search

[7]Notice that in general we can consider heterophilic bias, where $b_1 < q$ and $b_2 < 1 - q$, although this is not possible when biases are location driven. We rule out values of $\gamma < 1/2$ without loss of generality since they correspond to cases we already consider by relabeling the locations.

meetings, we get

$$\begin{aligned}\beta_1(1+r) &= b_1 r + \beta_1 b_1 + (1-\beta_2)(1-b_1)\\ \beta_2(1+r) &= b_2 r + \beta_2 b_2 + (1-\beta_1)(1-b_2)\end{aligned}$$

Solving this system of equations gives the steady-state network homophily values:

$$\beta_1 = \frac{b_1 r + 1 - b_2}{r + 1 - b_1 + 1 - b_2}; \qquad \beta_2 = \frac{b_2 r + 1 - b_1}{r + 1 - b_2 + 1 - b_1}. \tag{III.2}$$

Proposition 1 *Network homophily β_i is always strictly increasing in b_i and strictly decreasing in b_{-i}. Under a homophilic bias, it is strictly increasing in r and strictly less than the corresponding homophilic bias b_i.*

Proof. The results follow from examining the partial derivatives of equations (III.2). First, $\frac{\partial \beta_i}{\partial b_i} > 0$ if and only if $r(r+1+1-b_{-i})+1-b_{-i} > 0$, which holds since $r > 0$ and $b_{-i} < 1$. Next, $\frac{\partial \beta_i}{\partial b_{-i}} < 0$ if and only if $r + 1 - b_i > b_i r$, which holds as $r + 1 - b_i > r > b_1 r$. Now assume there is a homophilic bias, i.e., $b_1 > q$ and $b_2 > 1 - q$. Thus $b_1 + b_2 > 1$. We have $\frac{\partial \beta_i}{\partial r} > 0$ if and only if $b_1 + b_2 > 1$, and similarly $\beta_i < b_i$ if and only if $b_1 + b_2 > 1$, completing the proof. ■

Notice that Proposition 1 is true for arbitrary specifications of b_1 and b_2.

The intuition is as follows. As b_i increases, new g^i nodes form a higher proportion of their randomly met out-links with other g^i nodes. These parent nodes also have a higher proportion of g^i out-neighbors, and so β_i increases. As b_{-i} decreases, the random meeting process for g^i nodes is unaffected. Since some of the parent nodes will be from g^{-i}, and these nodes now have a higher proportion of g^i neighbors, g^i nodes will form a higher proportion of within-group links.

Now consider the case of homophilic bias. Since the resulting network homophily is increasing in r, *the meetings formed through the search process have a dampening effect on homophily.* This phenomenon is central to our analysis and deserves elaboration. When searching the network, a node typically has parent nodes of both types, and these parent nodes all have a local neighborhood that is homophically biased. Especially, parent nodes of the other type are

relatively more connected to nodes like themselves. Thus, the set of parents' neighbors is a mixture of neighborhoods, including some biased towards the other group, and therefore has a more neutral composition than the neighborhood of same-group nodes. In other words, meeting the friends of distinct friends opens up possibilities for diverse relationships and, as such, search acts as a mitigating force with respect to the bias in random meetings.

Introduce $\beta_{10} = \frac{1-b_2}{1-b_1+1-b_2}$ as the limiting (group 1) network homophily when r tends to zero, or $m_r \ll m_s$. It is the lowest possible value of network homophily for given values of q, b_1 and b_2.[8] When biases are location-driven, $\beta_{10} = q$ and hence a homophilic bias implies network homophily.[9] However, when b_1 and b_2 can vary freely, the dampening effect can be so strong as to overturn the homophilic bias, resulting for instance in $\beta_2 < 1 - q$.[10]

We next study *relative homophily* $H_1 = (\beta_1 - q)/(1 - q)$ and $H_2 = (\beta_2 - (1 - q))/q$. Relative homophily is positive when a group forms a higher proportion of its links within the group than would be implied by the population sizes, and is normalized to have a maximal value at unity. It provides a standard normalized index of homophily, see Coleman (1958) and Currarini, Jackson & Pin (2007). The following result shows how relative homophily changes as the composition of society varies.

Proposition 2 *H_1 is symmetric around $q = 1/2$. It is equal to zero at $q = 0$ and 1; it increases from $q = 0$ to $q = 1/2$, and decreases from $q = 1/2$ to $q = 1$, and is concave.*

Proof. Substituting from equation (II.1), we have

$$H_1(q) = \frac{(1-2\gamma)^2 rq(1-q)}{rq(1-q) + (1 + (1-2q)^2 r)\gamma(1-\gamma)}$$

From this expression, it is easily verified that $H_1(q) = H_1(1-q)$ and that $H_1(0) = H_1(1) = 0$.[11]

[8] We show below that β_{10} also plays a crucial role in degree distributions.

[9] More generally, the network exhibits homophily for all parameters if and only if $\beta_{10} = q$ and $\beta_{20} = 1 - q$.

[10] Interestingly, network heterophily can emerge for at most one of the two groups. This echoes Proposition 3 in Currarini, Jackson & Pin (2007).

[11] More generally, we can show that both properties hold provided that $\beta_{10} = q$.

The first derivative of H_1 is

$$\frac{\partial H_1(q)}{\partial q} = \frac{(1-2q)r(r+1)(2\gamma-1)^2\gamma(1-\gamma)}{\left((r+1)\gamma(1-\gamma)+(2\gamma-1)^2qr(1-q)\right)^2},$$

which has the same sign as $1-2q$, proving that H_1 is increasing below $q = 1/2$ and then decreasing. To show concavity, write the second derivative as

$$\frac{\partial^2 H_1(q)}{\partial q^2} = \frac{2\gamma(1-\gamma)(2\gamma-1)^2r(r+1) * (r(3q^2-3q+1)-\gamma(1-\gamma)(3r(2q-1)^2-1))}{-r^3(\gamma(1-\gamma)((2q-1)^2+1)+q(1-q))^3}.$$

The denominator is negative, and the term in the numerator before the asterisk is positive, so H_1 is concave if and only if $r(3q^2-3q+1)-\gamma(1-\gamma)(3r(2q-1)^2-1) > 0$. Dividing by r and rearranging, we must show that $\gamma(1-\gamma)(3(2q-1)^2-1/r)+3q(1-q) < 1$. $\gamma(1-\gamma) \leq 1/4$ and $-1/r \leq 0$; using these inequalities and collecting terms proves the result. ■

Thus, in the extreme cases where one group dominates society, relative homophily disappears. Natural mixing occuring at each location tends to homogenize meetings, and this effect overcomes the impact of location biases when sizes become very asymmetric. In all other cases, however, relative homophily is positive, and is strongest for intermediate size groups, reaching a maximum when the groups have equal size. Interestingly, the equilibrium mixing model of Currarini, Jackson & Pin (2008) generates an analogous result through a very different analysis, and their result is supported empirically looking at racial composition in the AddHealth data.

The next result shows how relative homophily responds to changes in the meeting process.

Proposition 3 *$H_1(q)$ is shifted up by an increase in r or by an increase in γ.*

Proof. The relevant derivatives are

$$\begin{aligned}\frac{\partial H_1}{\partial r} &= \frac{(1-2\gamma)^2\gamma(1-\gamma)q(1-q)}{(rq(1-q)+\gamma(1-\gamma)(1+(1-2q)^2)r)^2}\\ \frac{\partial H_1}{\partial \gamma} &= \frac{(2\gamma-1)q(1-q)r(1+r)}{(rq(1-q)+\gamma(1-\gamma)(1+(1-2q)^2)r)^2},\end{aligned}$$

both of which are easily verified as being positive. ■

That is, for a given society and homophilic biases, decreasing the role of search-based meetings increases relative homophily. This results from the "dampening effect" of search-based meetings on network homophily. The more prevalent are search-based meetings in the formation process, the lower homophily will be, as "friends of friends" are less likely to be of the same type than are individuals met through the biased random meetings. Since there is a homophilic bias whenever $\gamma > 1/2$, this result is consistent with Proposition 1 which describes network homophily more generally. Finally, increasing the location bias, all else equal, increases relative homophily for any value of r, since this is the parameter thats controls the extent to which random meetings are exogenously biased.

IV The relationship between local homophily and degree

Henceforth, we make use of the powerful mean-field approximation in analyzing the growing system of nodes and links. In this section, we want to analyze how the local homophily surrounding an individual varies with its degree. Since out-degree is homogenous, we keep track of in-degree, which varies substantially across individuals.

Consider a node $i \in g^1$. Let k_{11}^t denote the number of in-links of i with other nodes from group g^1 at time t, while k_{12}^t is the number of in-links with nodes from the other group. Then, $k_1^t = k_{11}^t + k_{12}^t$ is the total in-degree of node i at time t. We define *node i homophily* as the proportion of i's in-neighbors who belong to i's group. When node $i \in g^1$, this is equal to k_{11}^t/k_1^t. We first derive the expressions for degree growth for arbitrary specifications of biases b_1 and b_2. We then show that individual homophily generally decreases with degree, limiting to population frequencies for highest degree nodes, and obtain some sharper results under location-driven meetings.

Consider the individual entering at some time $t > i$. With what probability does this node form a link to i, and hence k_{11}^t or k_{12}^t increments?

$$P(k_{11}^{t+1} = k_{11}^t + 1) = qp(\frac{b_1 m_r}{qt} + \frac{m_s}{m m_r}(k_{11}^t \frac{b_1 m_r}{qt} + k_{12}^t \frac{(1-b_1)m_r}{(1-q)t}))$$

The new node t is in g^1 with probability q, and forms a link with i conditional on having met him with probability p. It meets i at random with probability $b_1 m_r$ (number of random meetings) divided by qt (number of nodes in g^1); or it meets i through search. In this case, one of i's in-neighbors must have been found randomly, which happens with probability $k_{11}^t \frac{b_1 m_r}{qt} + k_{12}^t \frac{(1-b_1)m_r}{(1-q)t}$. To see this, note that i has k_{11}^t out-neighbors in g^1, each of which is met at random by t with probability $\frac{b_1 m_r}{qt}$; similarly, i has k_{12}^t out-neighbors in g^2, each of which is met at random by t with probability $\frac{(1-b_1)m_r}{(1-q)t}$. Given that one of i's out-neighbors was met at random by t, i is met through search with probability $\frac{m_s}{m m_r}$, since t meets m_s nodes through search and has m_r parent nodes to search from, each of whom has on average m out-neighbors. By similar reasoning we get

$$P(k_{12}^{t+1} = k_{12}^t + 1) = (1-q)p(\frac{(1-b_2)m_r}{qt} + \frac{m_s}{m m_r}(k_{11}^t \frac{(1-b_2)m_r}{qt} + k_{12}^t \frac{b_2 m_r}{(1-q)t})).$$

Thus, under a continuous mean-field approximation, k_{11}^t and k_{12}^t satisfy the following system of differential equations with initial conditions $k_{11}^i = k_{12}^i = 0$

$$\begin{aligned}
\frac{\partial k_{11}^t}{\partial t} &= m\frac{r}{1+r}b_1\frac{1}{t} + \frac{1}{1+r}b_1\frac{k_{11}^t}{t} + \frac{1}{1+r}(1-b_1)\frac{q}{1-q}\frac{k_{12}^t}{t} \\
\frac{\partial k_{12}^t}{\partial t} &= m\frac{r}{1+r}(1-b_2)\frac{1-q}{q}\frac{1}{t} + \frac{1}{1+r}b_2\frac{k_{12}^t}{t} + \frac{1}{1+r}(1-b_2)\frac{1-q}{q}\frac{k_{11}^t}{t}
\end{aligned}$$

Solving these equation provides the expressions for degree growth, resulting in

Proposition 4

$$\begin{aligned}
k_{11} &= mr[\beta_{10}(\frac{t}{i})^{1/(1+r)} + (1-\beta_{10})(\frac{t}{i})^{(b_1+b_2-1)/(1+r)} - 1] \\
k_{12} &= mr\frac{1-q}{q}\beta_{10}[(\frac{t}{i})^{1/(1+r)} - (\frac{t}{i})^{(b_1+b_2-1)/(1+r)}].
\end{aligned}$$

Proof. See the Appendix. ■

To obtain the total in-degree of node i as a function of time, we simply sum the two equations of Proposition 4. This yields $k_1 = mr[\frac{\beta_{10}}{q}(\frac{t}{i})^{1/(1+r)} + (1-\frac{\beta_{10}}{q})(\frac{t}{i})^{(b_1+b_2-1)/(1+r)} - 1]$. We can now see how results of Jackson & Rogers (2007) are obtained as special cases. Two specific situations give us back the same expression as in the model without homophily. First if there is no homophilic bias. This corresponds to $b_1 = q$ and $b_2 = 1-q$, hence $\beta_{10} = q$ and $b_1 + b_2 = 1$ which leads to $k_1 = mr[(\frac{t}{i})^{1/(1+r)} - 1]$. Second if, in contrast, the homophilic bias is extreme and nodes do not meet any nodes from the other group. Then, $b_1 = b_2 = 1$ which also leads to $k_1 = mr[(\frac{t}{i})^{1/(1+r)} - 1]$. Thus, Proposition 4 indeed implies Theorem 1 in Jackson & Rogers (2008).[12]

This understanding of degree growth is necessary to study the relationship between degree and homophily.

Proposition 5 *There exist functions h^j, $j = 1, 2$, such that at any time t, the homophily of a node with degree k belonging to group g^j is equal to $h^j(k)$. Further, for all $k > 0, (h^j)'(k) < 0$ and $\lim_{k\to\infty} h^j(k) = q^j$.*

Proof. Without loss of generality, consider g^1. Introduce $\alpha_1 = 1/(1+r)$ and $\alpha_2 = (b_1+b_2-1)/(1+r) < \alpha_1$. Define the functions f and g over $[1, +\infty[$ as follows

$$\begin{aligned} f(x) &= mr[\frac{\beta_{10}}{q}x^{\alpha_1} + (1-\frac{\beta_{10}}{q})x^{\alpha_2} - 1] \\ g(x) &= \frac{\beta_{10}x^{\alpha_1} + (1-\beta_{10})x^{\alpha_2} - 1}{\frac{\beta_{10}}{q}x^{\alpha_1} + (1-\frac{\beta_{10}}{q})x^{\alpha_2} - 1} \end{aligned}$$

From Proposition 4, we know that $k_1(t) = f(t/i)$ and $k_{11}(t)/k_1(t) = g(t/i)$. Function f is increasing, hence admits an inverse f^{-1}. Define $h(x) = g(f^{-1}(x))$. We have: $k_{11}(t)/k_1(t) = h(k_1(t))$ and this relation can be expressed as a function of k_1, without explicit dependence on t and i. Next, compute the derivative of h: $h'(k) = (f^{-1})'(k)g'(f^{-1}(k))$. Here, $(f^{-1})' > 0$ and

[12]The two situations are of course not equivalent. For instance, we see that $k_{11} = qk_1$ when there is no homophilic bias while $k_{11} = k_1$ with an extreme homophilic bias.

$g'(x)$ has the same sign as

$$-\beta_{10}\frac{1-q}{q}(\alpha_1-\alpha_2)x^{\alpha_1+\alpha_2-1}[1-\frac{\alpha_1}{\alpha_1-\alpha_2}x^{-\alpha_2}+\frac{\alpha_2}{\alpha_1-\alpha_2}x^{-\alpha_1}]$$

The expression between brackets is increasing in x and equal to 0 when $x=1$. Thus, $g'(x)<0$ if $x>1$ and $h'(k)<0$ if $k>0$. In addition, $\lim_{x\to\infty} g(x)=q$. ■

Thus, nodes with higher degree have less homophilic immediate neighborhoods. The intuition for the result comes, again, from the differences between the two meeting processes. Larger degree nodes get a higher proportion of their links via the search process. Recall, meeting friends of friends opens up access to more diverse nodes. Conversely, nodes met through the network are also found by more diverse individuals. Thus, larger degree nodes get relatively more links from nodes of the other group. In the limit, for nodes with extremely high degree, this effect dominates the bias inherent in the random meetings.[13]

The way in which an individual's neighborhood composition limits to the population frequencies is non-trivial. Notice that if a particular individual became connected to a large proportion of others over time, then his neighborhood would necessarily approximate population frequencies. However, we emphasize that in our model, even though an individual's degree grows without bound, the proportion of others to whom he is connected still vanishes over time. This happens because the entry rate of new individuals is constant, while the probability for existing individuals to acquire a new link in any period goes to zero.

We now derive three additional properties under location-based random meetings.

Proposition 6 *Suppose that biases are location-driven. Then,*

$$h^j(k)=q^j+(1-q^j)\frac{(1+k/(mr))^{b_1+b_2-1}-1}{k/(mr)}.$$

For all $k>0$, $(h^j)''(k)>0$, $\partial h^j/\partial q^j(k)>0$ *and* $\partial^2 h^j/\partial q^j\partial k(k)>0$.

[13]Pin (2009) independently derives a similar limit result in the context of a more general model, and shows as well that the limit result does not obtain in the presence of additional sources of bias.

Proof. Analogous to the previous proof, define $f(x) = mr(x^{\alpha_1} - 1)$; hence $f^{-1}(k) = (1 + k/mr)^{1/\alpha_1}$. Given that $g(x) = q + (1-q)(x^{\alpha_2} - 1)/(x^{\alpha_1} - 1)$ and $\alpha_2/\alpha_1 = b_1 + b_2 - 1$, we obtain the expression for $h(k) = g(f^{-1}(k))$. Without loss of generality, we can set $mr = 1$ in what follows. Introduce $b = b_1 + b_2 - 1$ and $y = 1 + k$. Next, $h'' = [(f^{-1})']^2 g'' \circ f^{-1} + (f^{-1})'' g' \circ f^{-1}$. Developing and substituting shows that h'' has the same sign as

$$\varphi(y) = y^{b+2}(1-b)(2-b) + y^{b+1}2b(2-b) - y^b b(1-b) - 2y^2$$

A detailed study of φ and its first three derivatives then shows that $\varphi(y) > 0$ if $y > 1$, hence that $h''(k) > 0$ if $k > 1$.

The explicit expressions for the derivative with respect to q are not trivial given that b is a function of q. We have

$$\begin{aligned} \frac{\partial h}{\partial k} &= -(1-q)\frac{(1+k)^{b-1}(1+(1-b)k) - 1}{k^2} \\ k^2(1+k)^{1-b}\frac{\partial^2 h}{\partial k \partial q} &= \psi(k) = 1 + (1-b)k - (1+k)^{1-b} + (1-q)(-\frac{\partial b}{\partial q})\left[\ln(1+k)(1+(1-b)k) - k\right] \end{aligned}$$

Also, note that $h(0) = q + (1-q)b$. Thus, $\frac{\partial h}{\partial q}(0) = 1 - b + \frac{\partial b}{\partial q}(1-q)$. We have: $b = 1 - \frac{\gamma(1-\gamma)}{q(1-q)(2\gamma-1)^2 + \gamma(1-\gamma)}$ and $\frac{\partial b}{\partial q} = -\frac{(2q-1)(2\gamma-1)^2\gamma(1-\gamma)}{[q(1-q)(2\gamma-1)^2 + \gamma(1-\gamma)]^2}$. Developing, we get that $\frac{\partial h}{\partial q}(0)$ has the same sign as $1 - (2\gamma-1)^2 q(2-q) - 3\gamma(1-\gamma) \geq 1 - (2\gamma-1)^2 - 3\gamma(1-\gamma) = \gamma(1-\gamma) > 0$ where the first inequality comes from the fact that $q(2-q) \leq 1$. Thus, $\frac{\partial h}{\partial q}(0) > 0$ and $1 - b > (-\frac{\partial b}{\partial q})(1-q)$.

Next, derive the function ψ with respect to k. We have:

$\psi'(k) = (1-b)(1-(1+k)^{-b}) + (1-q)(-\frac{\partial b}{\partial q})\left[(1-b)\ln(1+k) - b(1-\frac{1}{1+k})\right]$ and

$(1+k)\psi''(k) = (1-q)(-\frac{\partial b}{\partial q})(1-b) - (1-q)(-\frac{\partial b}{\partial q})b(1+k)^{-1} + b(1-b)(1+k)^{-b}$

Here, $\psi''(0) = b(1-b) + (1-q)(-\frac{\partial b}{\partial q})(1-2b)$. Since $1 - b > (-\frac{\partial b}{\partial q})(1-q)$, we have $\psi''(0) \geq (1-q)(-\frac{\partial b}{\partial q})(1-b) > 0$. Also, $\lim_{k\to\infty} \psi''(k) = 0^+$. By looking at its derivative, we see that the function $(1+k)\psi''(k)$ is either decreasing, or increasing and decreasing. In either case, since it is positive when $k = 0$ and when $k \to \infty$, it must be greater than or equal to zero for any k. Thus, $\psi'' > 0$ if $k > 0$ hence ψ' is increasing. Since $\psi'(0) = 0$, $\psi' > 0$ if $k > 0$. Thus, ψ

is increasing and as $\psi(0) = 0$, $\psi > 0$ and $\frac{\partial^2 h}{\partial k \partial q} > 0$ if $k > 0$. Finally, given that $\frac{\partial h}{\partial q}(0) > 0$ and $\frac{\partial h}{\partial q}$ is increasing in k, $\frac{\partial h}{\partial q} > 0$, $\forall k$. ■

As could be expected, homophily decreases with degree at a decreasing rate. In fact, this is true, and the functional form given in the proposition is valid, so long as $\beta_{10} = q$, which encompasses the case where of location-driven biases. Homophily is also higher in larger groups. Members of a relatively larger group, all else equal, have a higher proportion of intra-group links due to a direct "size effect." Perhaps more surprisingly, the relationship between homophily and degree is flatter for individuals in larger groups. In other words, the positive effect of group size on homophily is greater for nodes with higher degrees. Thus, the difference in homophily between low-degree and high-degree nodes is smaller in larger groups. Overall these results provide four sharp empirical predictions. In the next section, we test these predictions on data from friendship networks among adolescents.

Simulations of network formation

In light of the fact that our theoretical analysis relies on a mean-field approximation, we present a set of computer simulations of the stochastic network formation process. This allows us to check directly the validity of the predictions against the results of the simulations.

Each simulation is run under location-driven biases for $T = 3000$ nodes with $m_r = 10$. We vary the parameters b_1, b_2, q, m_s, and p. Notice that this generates variation in γ, since its value is determined by b_1, b_2, and q when biases are location-driven. This provides enough variation in the formation process to have confidence in the global predictions of the model. [14]

The parameters are chosen such that $b_i m_r$ and m_s take integer values. This is important, since in the discrete process, there is not an obvious interpretation of non-integer meeting parameters. Taking these parameters as given, the possibility arises that a given node may not be able to meet as many individuals as required, due to network-dependent constraints. In fact, this is bound to happen for the oldest nodes, when $t \leq b_i m_r$. In these cases, we specify the process so

[14] Specifically, we take values of $b_2 m_r$ from 5 to 9, values of $b_1 m_r$ from $b_2 m_r$ to 9, values of m_s of 2, 6, 10, and 30, and values of p of 0.6 and 1. This allows us to solve for values of q and γ in each case via equations (II.1). Since $b_1 \geq b_2$, we always have $q \geq \frac{1}{2}$.

that node t meets as many nodes as possible; in this case, that would be all $t-1$ existing nodes. In all of the simulations we conducted, no network constraints bind beyond roughly $t = 30$.

One quantity that is easy to measure for each simulation is the resulting network homophilies β_1 and β_2. Over the 120 combinations of parameters we studied, the average absolute difference between the predicted and realized value of β_1 was 0.0075 with an average value of 0.720, and for β_2 was 0.0099 with an average value of 0.550. There appears to be a slight positive bias in our predictions. However, we conducted a smaller set of simulations with values of T as large as $10,000$ which show that such bias is the result of our finite simulations, and appears to vanish as T grows.

There are many ways to measure the quality of the theoretical predictions with the simulated data. Given that the results of Section IV are most central to our motivating questions, we focus on discussing the relationship between degree and individual homophily in the simulated networks.[15] In summary, we find strong support for the model's predictions. The support is exemplified in Figure 1, which depicts the relationship between individual homophily and degree for a typical simulation. Each point represents the mean individual homophily among all nodes of a particular in-degree. The bounds from Proposition 6 are depicted as horizontal lines in the the figure. The upper bound is given by $h(0) = q + (1-q)(b_1 + b_2 - 1)$, and the lower bound is given by $\lim_{k\to\infty} h(k) = q$. Values cluster very near the upper value for low degree nodes, and appear to asymptote to the lower value, exhibiting a decreasing, convex shape as a function of degree, as predicted.

V Relationships among the distributions of links

Even with only two groups, the distribution of links becomes substantially more complex as nodes can connect to both same and different type nodes, and one wants to keep track of the different kinds and sources of links. In this context, we can keep track of *seven* different degree distributions rather than one. Define F_{ij} as the distribution of the in-degrees of type g^i nodes

[15] A more complete set of results from the simulations is available upon request.

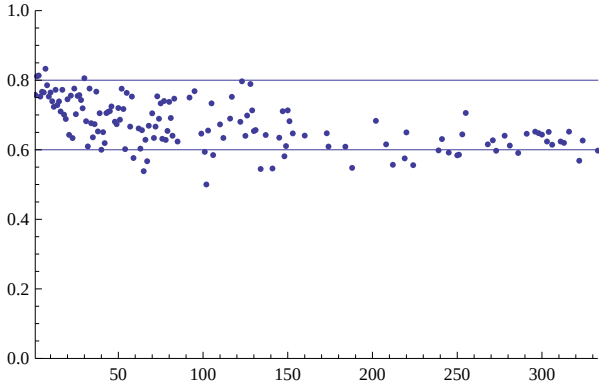

Figure 1: *Individual homophily as a function of in-degree for a typical simulation. The parameters here are* $q = 0.6$, $b_1 = .8$, $b_2 = .7$, $r = \frac{1}{3}$ *and* $p = 1$. *The implied value of* γ *is approximately* 0.86.

paying attention only to links coming from nodes of type g^j, $i, j = 1, 2$. Then F_1 and F_2 are the standard in-degree distributions of g^1 and g^2 nodes (ignoring the types of in-neighbors), and finally F is the total in-degree distribution of the entire society.

We emphasize that we are able to derive many of the results of this section for arbitrary specifications of biases b_1 and b_2. It is not until we get to Proposition 10 that we specialize to consider biases driven by γ and location choices.

To obtain, for example, F_1, observe that $F_1(k) = 1 - i/t$, where t is an arbitrary time period and i is the node that has in-degree k at time t under the mean-field system. Thus $t/i = 1/(1 - F_1(k))$. This defines F_1 implicitly as a function of k, and a similar method works for the other distributions. While these equations do not usually yield closed-form solutions, they still allow us to derive important properties of the degree distributions. Our first such result orders the degree distributions as the number of out-links is varied.

Proposition 7 *Fix* q, b_1, b_2 *and* r. *Let* F_{ij} *be the distributions corresponding to the parameter* m *and let* F'_{ij} *be the distributions corresponding to* m'. *If* $m' > m$, *then* F'_{ij} *strictly first-order stochastically dominates* F_{ij}, *for every* $i, j = 1, 2$.

Proof. It is enough to show that the expressions for the k_{ij} are increasing in m, which follows directly from Proposition 4. We can then apply Lemma 1 (see the Appendix) to prove the result.

■

This implies the result in Theorem 7 of Jackson & Rogers (2007), where the case of only one group is considered. Next we turn to approximating the upper tail of the various degree distributions. It turns out that all of the degree distributions have similar upper tails; in particular, they all tend to a scale-free distribution with exponent $-(1+r)$.

Proposition 8 *When k is large,*

$$\begin{aligned}
\ln(1-F_{11}(k)) &\sim \ln(mr\beta_{10}) - (1+r)\ln(k) \\
\ln(1-F_{12}(k)) &\sim \ln(mr\beta_{10}\frac{1-q}{q}) - (1+r)\ln(k) \\
\ln(1-F_{1}(k)) &\sim \ln(mr\beta_{10}\frac{1}{q}) - (1+r)\ln(k)
\end{aligned}$$

Proof. From Proposition 4, we can obtain the implicit equation defining $F_{11}(k)$:

$$k = mr[\beta_{10}(1-F)^{-1/(1+r)} + (1-\beta_{10})(1-F)^{-(b_1+b_2-1)/(1+r)} - 1],$$

which we write as

$$k = mr\beta_{10}(1-F)^{-1/(1+r)}[1+\varepsilon(k)],$$

where $\varepsilon(k) = \frac{1-\beta_{10}}{\beta_{10}}(1-F)^{(2-b_1-b_2)/(1+r)} - (1-F)^{1/(1+r)}$. Since $b_1+b_2 < 2$, we have $\lim_{k\to\infty}\varepsilon(k) = 0$. Taking the log we get:

$$\ln(1-F) = (1+r)\ln(mr\beta_{10}) - (1+r)\ln k + \eta(k),$$

where $\eta(k) = (1+r)\ln(1+\varepsilon(k))$, and so $\eta(k)$ tends to zero as k tends to infinity. Similar reasoning works for the other distributions. ■

This result says that all seven distributions have a thick upper tail following a power law with exponent $-(1+r)$. In other words, the relative proportions of links coming to nodes of either type from nodes of either type are all identical for sufficiently high degree. This provides

a particularly sharp and empirically testable prediction of the model. The upper tails of the distributions for nodes in group g^2 can be derived analogously.

Next, we ask how F_{i1} and F_{i2} compare to each other. That is, we focus on one group g^i, and compare the distributions for that group of in-degrees coming from each of the two groups. We find that the answer very much depends on the size of group i.

Proposition 9 *Fix $q > 1/2$. Then*

(i) F_{11} FOSD F_{12};

(ii) If $b_1 + b_2 < 2$, then F_{22} never FOSD F_{21};

(iii) F_{21} FOSD F_{22} if and only if $(2q-1)(1-b_1) \geq (1-q)(b_1+b_2-1)$.

Proof. For (i), use Proposition 4 to write

$$\begin{aligned} k_{11} &= mr\beta_{10}\left[(\frac{t}{i})^{1/(1+r)} - (\frac{t}{i})^{(b_1+b_2-1)/(1+r)}\right] + mr\left[(\frac{t}{i})^{(b_1+b_2-1)/(1+r)} - 1\right] \\ k_{12} &= mr\beta_{10}\left(\frac{1-q}{q}\right)\left[(\frac{t}{i})^{1/(1+r)} - (\frac{t}{i})^{(b_1+b_2-1)/(1+r)}\right]. \end{aligned}$$

Given that $q > 1/2$ and that $b_1 + b_2 \geq 1$, we know that $\frac{1-q}{q} < 1$ and the second term in the first equation is non-negative. Thus $k_{11}(t/i) > k_{12}(t/i)$ for all $t \geq i$, which allows us to apply Lemma 1.

Now consider the expressions for k_{22} and k_{21} obtained from the above equations by switching the group labels 1 and 2 and switching q with $1-q$. When $q > 1/2$ (meaning g^1 is the majority group) then $\frac{q}{1-q} > 1$, and when $b_1 + b_2 < 2$ (meaning there is at least some inter-group linking) then for large values of t/i the second term in the expression for k_{22} becomes negligible, in which case $k_{22} < k_{21}$ in the upper tail, proving (ii) by application of Lemma 1.

For (iii), introduce the function $\psi(x) = q\beta_{20}[x^{\alpha_1} - x^{\alpha_2}] - (1-q)[\beta_{20}x^{\alpha_1} + (1-\beta_{20})x^{\alpha_2} - 1]$, where $\alpha_1 = \frac{1}{1+r}$ and $\alpha_2 = \frac{b_1+b_2-1}{1+r}$. Note that $\psi(x) \geq 0$ if and only if $k_{21}(x) \geq k_{22}(x)$. Observe that $\psi(1) = 0$. Also,

$$\psi'(x) = x^{\alpha_1 - 1}[(2q-1)\beta_{20}\alpha_1 - (1-q+(2q-1)\beta_{20})\alpha_2 x^{\alpha_2 - \alpha_1}]$$

Since $\alpha_1 \geq \alpha_2$, the second term of the RHS is weakly increasing in x. There are two cases. First, $\psi'(1) \geq 0$, in which case $\forall x \geq 1, \psi'(x) \geq 0$, thus ψ is weakly increasing and $\forall x \geq 1, \psi(x) \geq 0$. Otherwise $\psi'(1) < 0$, in which case ψ' is first negative then positive above 1 (since $\psi'(\infty) = \infty$), hence ψ is first decreasing and then increasing, which also means that ψ is first negative and then positive above 1. Therefore, F_{21} FOSD F_{22} if and only if $\psi'(1) \geq 0$. The condition reduces to

$$(2q-1)\beta_{20}\alpha_1 \geq [1-q+(2q-1)\beta_{20}]\alpha_2$$

which after some algebra can be written as

$$(2q-1)(1-b_1) \geq (1-q)(b_1+b_2-1)$$

■

These results express the interplay of two effects. On the one hand, there is a direct size effect through which nodes receive more links from the larger group. On the other hand, homophily leads nodes to receive relatively more links from nodes from the same group. In the larger group, both effects are aligned which implies that F_{11} FOSD F_{12}. In the smaller group, however, these effects pull in opposite directions. The third item in the proposition says that if homophily is not too large, the size effect dominates and F_{21} FOSD F_{22}. In contrast, the second item says that even if homophily is very large, as long as is is not perfect ($b_1 = b_2 = 1$), the homophily effect cannot dominate the size effect. The explanation lies with nodes of high degree. We can use our previous approximation result to show that, in the tails, F_{22} always lies above F_{21}. In other words, the size effect dominates for the hubs of the smaller group, and they tend to get relatively more connections from nodes of the larger group. This is related to the fact that the largest degree nodes have the least homophilic set of neighbors. In other words, the hubs in the minority group have the greatest proportion of their in-neighbors from the majority group.

Now let us return to the original model in which the biases b_1 and b_2 are driven by location choices, γ. In this case, we are able to derive sharper predictions concerning the relationships among the various degree distributions, as detailed in the remaining results of this section.

Working under the original model, previous equations reduce to $k_{11} = mr[q(\frac{t}{i})^{1/(1+r)} + (1-q)(\frac{t}{i})^{(b_1+b_2-1)/(1+r)} - 1]$, $k_{12} = mr(1-q)[(\frac{t}{i})^{1/(1+r)} - (\frac{t}{i})^{(b_1+b_2-1)/(1+r)}]$ and $k_1 = mr[(\frac{t}{i})^{1/(1+r)} - 1]$. This gives us the following striking result.

Proposition 10 *When biases are location-driven, $F_1 = F_2$.*

This provides a particularly strong empirical prediction, as independent of the homophilic biases, the relative group sizes, and the proportion of links formed through the random meeting process, the in-degree distributions of the two groups must be identical.

We can also specialize the previous results on first order stochastic dominance to achieve the following.

Corollary 1 *When biases are location-driven, F_{21} FOSD F_{22} if and only if $\gamma(1-\gamma) \geq q * (1-q)/(1+2(1-q)(2q-1))$*

Proof. The result follows from part (iii) of Proposition 9 by substituting from equation (II.1). ■

This condition requires that γ be lower than or equal to some threshold value. Also, we can see that this threshold is increasing in q. As the size of the larger group increases, the size effect becomes relatively more important and F_{21} ends up dominating F_{22} for a larger range of the parameters.

The final result in this section describes how the distributions of inter- and intra-group links respond to changes in the homophilic bias.

Proposition 11 *Assume biases are location-driven. Fix q, m and r and take $\gamma < \gamma'$. Let F_{ij} be the distributions corresponding to γ and let F'_{ij} be the distributions corresponding to γ', for*

$i, j = 1, 2$. *Then* F'_{11} *and* F'_{22} *strictly FOSD* F_{11} *and* F_{22}*, while* F_{21} *and* F_{12} *strictly FOSD* F'_{21} *and* F'_{12}.

Proof. Observe that $b_1 + b_2 - 1$ increases with γ. This means that $k'_{11}(x) \geq k_{11}(x)$ and $k'_{22}(x) \geq k_{22}(x)$ while $k'_{12}(x) \leq k_{12}(x)$ and $k'_{21}(x) \leq k_{21}(x)$. The result then follows from Lemma 1. ■

When the homophilic bias increases, no matter the group sizes, individuals tend to form more links within their own groups, and fewer links across groups. Notice that in the presence of a utility function over links and the compositions of neighbors, one can use these results on degree distributions to begin to conduct a welfare analysis. This is the issue we turn to in the following section.

VI Welfare implications

The framework studied here provides a vehicle with which to illustrate a range of welfare implications for individuals as a function of the parameters that describe the formation of the network in which they live. To that end we posit that agents in group i have preferences over the size and composition of their neighborhoods of the form

$$u_i(k_i, k_{-i}) = k_i^{\beta} + \alpha k_{-i}^{\beta},$$

for $0 < \alpha$ and $0 < \beta < 1$, where k_i and k_{-i} are the number of inlinks from the individual's own and other group, respectively. That is, utility is separable, and increasing and concave in connections from either group. The marginal utility from an own-group connection is higher (lower) than that from an other-group connection when α is smaller (larger) than one.[16] Notice that these preferences are consistent with the network formation model we have analyzed. We have in mind that any given agent is compatible with another agent with probability p, and

[16] Of course, the conclusions derived in this section are sensitive to the utility specification. We aim only to illustrate how welfare analysis can be addressed in our framework, focusing on preferences that seem natural in this context.

adding a connection to a compatible agent increases one's utility according to u_i, depending on the types of the agents.

The first question we address is how preferences over friendships induce preferences over the extent of bias in the network formation process. Conditional on a particular number of total inlinks, individuals have an optimal proportion of own-group to other-group friends, and this induces an optimal level of bias. In particular we can show that, while the underlying preferences over friendships of the two groups are clearly symmetric, *the minority group prefers to have more bias in random meetings* (γ) than the majority group. This is because the size effect makes it relatively less likely that a minority group individual randomly meets others in his own group. Moreover, the difference in preferences between the groups is increasing in the difference in their relative size.

Figure 2 plots average utility for individuals in the majority group (solid) and minority group (dahsed) as a function of γ. Notice first that both utility functions are single-peaked, meaning that each group has an optimal level of γ, since that determines the average composition of neighborhoods for individuals in each group. Next, for low amounts of bias the average utility in the majority group is higher, whereas for high amounts of bias, individuals in the minority group have higher average utility. Finally, the optimal level of bias is clearly greater for the minority group, since more bias is necessary for these individuals to reach their optimal level of local homophily.

We now explore how the ratio of random meetings to search-based meetings (r) effects average utility. There are two channels through which r affects utility. The first, as shown by Jackson and Rogers (2007) is that the degree distributions F_1 and F_2 change according to mean-preserving spreads (MPS) as r decreases. When $\beta < 1$, utility is concave in degree, and will thus average utility tends to increase with r. The second channel is that r affects the level of homophily: as r increases, the dampening effect of search is mitigated and local homophily tends to be higher. Whether this effect is positive or negative depends on how local homophily compares to the optimal level according to preferences.

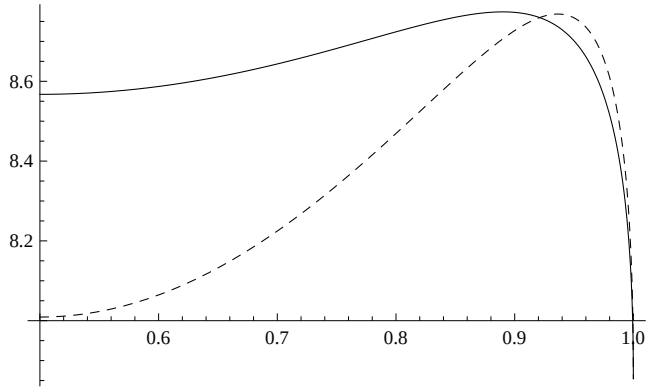

Figure 2: *Average utility in the majority group (solid) and minority group (dashed) as a function of γ. The qualitative features are representative of a wide range of parameters, and are depicted for $\alpha = \beta = .5$, $r = 1$, $q = .6$, and $m = 100$.*

It turns out that, typically, the MPS effect dominates the homophily effect, and thus average utility increases with r. Figure 3 (left panel) depicts a typical scenario. In order to better illustrate the specific impacts of homophily on welfare, we next report results for atypical regions of the utility parameter space. The homophily effect may overcome the MPS effect when two conditions are met. First, β must be close to one so that the effect of concavity is small and, second, it must be valuable to have friends from the other group, so that $\alpha < 1$. As depicted in Figure 3 (right panel), such parameters can result in a non-monotone effect of r on average utility. In particular, notice that for sufficiently high value of r, utility decreases when r is increased, as more random meetings increases local homophily, pulling the mixture of friends further away from the optimal level.

The combination of nearly linear utility and a preference for other-group friends can create a second interesting effect. Typically, the marginal utility of an additional friend tends to decrease over time, as additional friends add less utility. But in this case marginal utility can be non-monotonic in degree. As depicted in Figure 4, marginal utility is increasing over an interior interval of degrees. This is the case even though local homophily is strictly decreasing in degree. When the value of other-group friends is high enough (and β is close to one) then as degree increases, the chance that the next link comes from the other group is increasing, and this creates the possibility of increasing marginal utility.

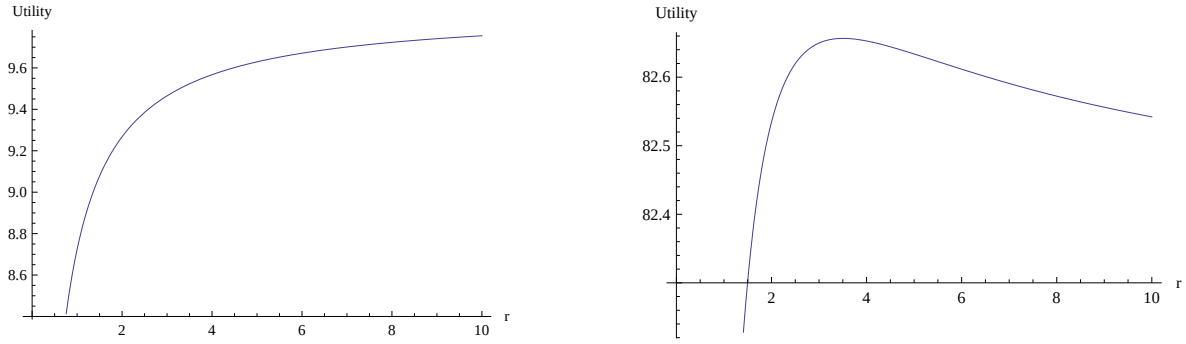

Figure 3: *Average utility in the majority group as a function of* r*. Left panel: the mean-preserving spread effect of* r *tends to make average utility increase. The chosen parameters are* $\alpha = \beta = .5$*,* $\gamma = .8$*,* $q = .6$*, and* $m = 100$*. Right panel: other-group friends are more valuable than own-group friends, and the effect on utility is non-monotone. For high values of* r*, the homophily-increasing effects of more random meetings cause average utility to decrease. The chosen parameters are* $\alpha = 2$*,* $\beta = .9$*,* $\gamma = .8$*,* $q = .6$*, and* $m = 100$*.*

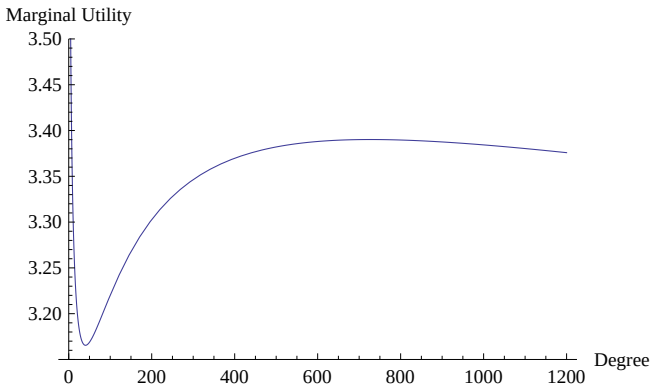

Figure 4: *Marginal utility in the majority group as a function of degree. The relationship is non-monotone and, in particular, increasing for an interior range of degree. The chosen parameters are* $\alpha = 20$*,* $\beta = .9$*,* $\gamma = .9$*,* $q = .6$*,* $r = 1$*, and* $m = 100$*.*

While we have illustrated only a few simple observations, the scope for conducting more detailed welfare comparisons is large. Further work could leverage our results that the degree distributions obey stochastic dominance shifts, and explore the consequences of other utility specifications.

VII Empirical analysis

In this section, we study how the predictions of the model compare with empirical properties of social networks. To do so, we analyze the AddHealth data, which contains extensive information on friendship networks in a sample of high schools in the U.S. This data allow us to look at homophily patterns in friendship networks with respect to gender composition. We first describe the data, as it applies to our analysis, in more detail. We then study the relationship between homophily and degree, checking whether the theoretical results of Propositions 5 and 6 hold empirically. Finally, we discuss other empirical properties in light of the model's predictions. Overall, the theoretical predictions of the model are remarkably well supported by this data.

The analysis below is based on the AddHealth study. We use the first wave of the in-school survey, which was conducted between September 1994 and April 1995 in a representative sample of American high schools. One objective of this survey was to collect information on *all* students within any particular school. Most relevant from our point of view, students were asked to name their best friends. They could name up to five male friends and five female friends. Friendship nominations are not necessarily reciprocal so these friendship networks are, as such, directed. To interpret this data through the lense of our model, we assume that friendships form over time, do not break, and that there is both a random and a search component to the meeting process.

In what follows, we focus on the in-degree and homophily of a student. A student's in-degree is defined as the number of other students (in the same school) who name her as one of their best friends. A student's homophily is the proportion of students with the same gender among all the students who name her as one of their best friends. It is well-defined only if the student receives at least one nomination. Overall, there are 142 schools and 67,916 students who receive at least one friendship nomination in our sample. The overall proportion of boys is 0.498. Overall network homophily is 0.577 for boys and 0.617 for girls which indicates that, indeed, students are relatively more likely to be friends with others of the same gender.[17] Very few links originating

[17]These figures also indicate a possible asymmetry between boys and girls. However, since these are aggregate figures, it leaves open the question of whether the asymmetry can be explained by school-level variations. We explore this below.

in one school of the sample end up in another school. So for our purposes, it is best to view the overall network as the collection of 142 smaller disconnected networks. The identification of our main regression below relies on two sources of variation. First, we have some variation across schools in terms of gender composition. Second, we have natural variation in degrees of students within schools. Figure 5 depicts the frequency distribution of in-degree in the data.

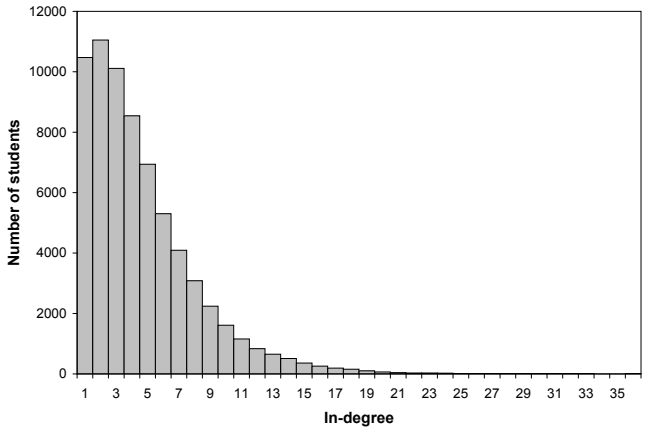

Figure 5: *The distribution of in-degree across students in the AddHealth sample.*

We now look at the relationship between homophily and degree at the individual level. Figure 6 depicts how average individual homophily varies with in-degree over the whole sample.[18] The relation clearly shows a decreasing and convex shape. The effect of degree also appears qualitatively important. The average individual homophily is equal to 0.748 among students with in-degree 1, 0.515 among students with in-degree 10, and 0.427 among students with in-degree 20. Of course, this evidence is only suggestive as the effect of gender composition (q) is not controlled for. To provide a rigorous test of our results, we conduct a set of linear regressions. Denote by h_{ij} the individual homophily of student i in school j, by k_{ij} the in-degree of student i in the friendship network of high school j, and by q_{ij} the proportion of students in high school j who have the same gender as i.[19] Our analysis in section IV shows that under the Location-Bias

[18]Dashed lines represent the 95% confidence interval.

[19]Thus, if i_1 and i_2 have the same gender and are in the same school, $q_{i_1 j} = q_{i_2 j}$ and if i_1 is a boy and i_2 is a

Table 1: Regression estimates of individual homophily.

	Whole sample	Boys	Girls
degree	-0.0528	-0.0558	-0.0452
	(0.0024)	(0.0026)	(0.0067)
degree^2	0.00059	0.00072	0.00058
	$(4 \cdot 10^{-5})$	$(6 \cdot 10^{-5})$	$(6 \cdot 10^{-5})$
q^j	0.53	0.567	0.268
	(0.018)	(0.017)	(0.087)
degree$\cdot q^j$	0.043	0.048	0.024
	(0.004)	(0.004)	(0.013)
Constant	0.512	0.456	0.689
	(0.011)	(0.012)	(0.045)
Observations	67916	32876	35040
R^2	0.091	0.10	0.089

Note: Robust standard errors are given in parentheses.

model individual homophily h_{ij} should depend on k_{ij} and q_{ij}. In particular, the relationship should be decreasing and convex in k_{ij}, increasing in q_{ij} and with a positive cross-derivative between k_{ij} and q_{ij}. In order to test these hypotheses, we estimate a linear regression of h_{ij} on k_{ij}, k_{ij}^2, q_{ij}, and $k_{ij}q_{ij}$. We run these regressions over the whole sample as well as separately for boys and girls. The estimation results are reported in Table 1.

The first column reports results for the whole sample. Remarkably, all coefficients have the predicted sign and are statistically significant at the 1% level. At $q = 0.5$, the estimated relationship between homophily and degree is decreasing and convex up through an in-degree of $k = 26$, which almost entirely covers the data's support.[20] The effects of q and of kq are also both clearly positive. So holding degree constant, homophily is larger in relatively larger groups. Finally, degree has a lower effect (in absolute value) on homophily in relatively larger groups. The second column reports results for boys only and, likewise, the third column for girls only. Again, all the coefficients have the predicted sign. They are also all statistically significant at the 1% level, except for the effect of kq for girls which is statistically significant at the 10% level.

girl of the same school, $q_{i_1 j} = 1 - q_{i_2 j}$.

[20]The proportion of students with degree 16 or lower is 99.2% and is 99.9% for those with degree 26 or lower.

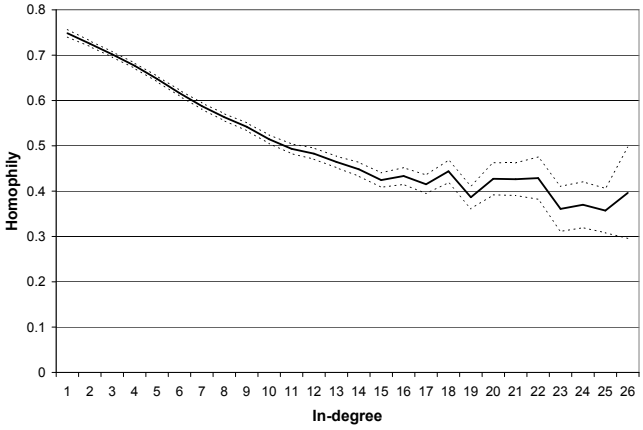

Figure 6: *The relationship between individual homophily and in-degree shows a decreasing and convex pattern.*

At $q = 0.5$, the predicted relationship between homophily and degree is decreasing and convex for degrees up to $k = 21$ for boys and $k = 28$ for girls. The effects of k and k^2 are quantitatively similar for boys and girls. However, the effects of q and kq are roughly twice as large for boys than for girls. This indicates some asymmetry between the link formation processes of boys and girls, which is not accounted for in the model.[21] In summary, we find strong empirical support for our theoretical predictions relating individual homophily to other network features.

We next use the closed form describing the homophily-degree relationship under the location-bias model to estimate the structural parameters of the network formation process, r and γ. The results are summarized in Table 2. We find that the bias parameter is roughly 0.7, and there are slightly more links formed through network-based search than through random meetings. The estimates of γ are tighter, reflecting the fact that the likelihood surface is flatter in the dimension of r. For this reason, the parameter differences between boys and girls are not significant.

Notice that there are two mechanisms in the model to generate higher network homophily. The first is to increase the location bias parameter. The second is to have more links formed

[21]There are a number of ways to introduce heterogeneity between the groups beyond their relative sizes. For instance, one could consider indexing the parameters of link formation such as m_r and m_s, by group identity.

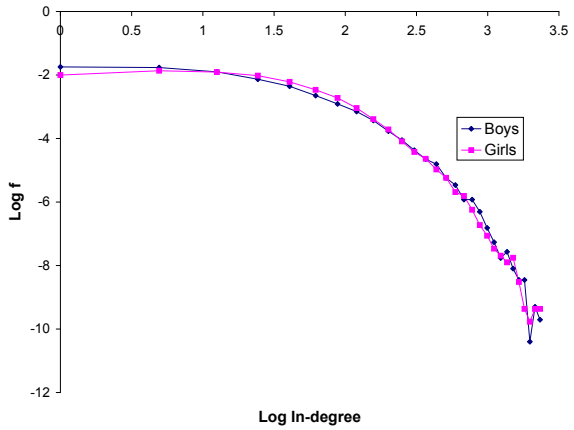

Figure 7: *The pdf of in-degrees for boys (blue) and girls (pink) shown in log-log scale.*

Table 2: Structural estimates of location-bias parameters.

	Whole sample	Boys	Girls
γ	0.698(0.015)	0.662(0.080)	0.729(0.014)
r	0.647(0.363)	0.138(0.302)	1.081(0.667)

through random meetings. If one observes only network homophily, then it would not be possible to distinguish these mechanisms. However, as demonstrated by these regressions, if one observes as well individual-level homophily, then it is possible to separate the effects of the two sources, since they imply different relationships between homophily and degree at the individual level.

The model yields further testable implications. In section V, we derive several results constraining the relationships among various degree distributions. However, the AddHealth data consists of a collection of 142 relatively small networks. This prevents an accurate estimation of the full degree distributions for each network, especially considering that the theoretical predictions rely on a mean-field approximation which becomes valid only for large networks. Still, under the Location-Bias model, F_1 should be equal to F_2 and should be independent of q (Proposition 10). So the distribution of in-degree for boys should be equal to the distribution

of in-degree for girls over the whole sample, provided the other parameters of link formation are fixed throughout the sample. Figure 7 depicts the pdf's of the two distributions in a log-log plot. While the two distributions are not precisely identical, they are reasonably close.[22]

Finally, in section III, we show that the relationship between relative homophily and school gender composition should have an inverse-U shape.[23] To check for this relationship, we regress H_j on q_j and q_j^2 over all schools. While we effectively obtain an inverse-U relationship, the coefficients are statistically not significant.[24] So the empirical relation between network homophily and gender composition is essentially flat over our sample. Note, however, that empirical values of q_j lie mostly between 0.4 and 0.6.[25] A flat relationship between homophily and gender composition around $q = 0.5$ is consistent with Proposition 2. However, the range is probably too small to lead to an informative test of the result.

VIII A general framework for biases

Our model allows for a relatively clean analysis of how homophily and network structure are affected by biases in the meeting process. However, biases in random meetings are only one way in which homophily might be generated at the individual-level. We now turn to discussing other potential sources of biases and their impact on network structure. One theme that emerges is that different kinds of biases have different implications for network structure, and thus can be empirically identified in the context of our framework. In particular, the relationship between an individual's degree and the local homophily in his neighborhood depend crucially on the nature of homophilic biases in the formation process.

One can think of our random meeting biases as the manifestation of institutionally driven

[22]The proportion of students with a very low in-degree (0 or 1) is somewhat higher for boys than for girls. Extending our model to account for such asymmetries, as in footnote 21, provides an interesting direction for future research; we return to this issue in the conclusion.

[23]Recall that Currarini, Jackson and Pin (2008) find empirical evidence of this inverse-U shape on the same data set, but with respect to racially-based homophily.

[24]We obtain $H_j = 0.053 + 0.479q_j - 0.400q_j^2$, with standard errors 0.110, 0.388, and 0.367, respectively. Notice that the graph of this curve is nearly flat over the range of q observed in the sample.

[25]Only 7 schools representing less than 4% of the students have a proportion of boys outside this range.

asymmetries in social structure. That is, under a homophilic bias in our model, for whatever reason, it is the case that individuals happen to meet similar individuals at a rate higher than what is suggested by their representation in the population. Conditional on who is found in the random meetings, an individual does his network-based search, and then forms his links in a completely unbiased manner, i.e., ignoring group labels. In this way, one can view the results of our model as implications of network structure generated by institutional constraints, such as neighborhood demographics sorting on the dimension of income or race.

In some contexts, additional biases driven by individual's preferences may play an important role as well. It is our purpose here to model such biases and discuss their impact on observed network patterns.

The most obvious manifestation of preference-based biases would be expressed in the probability of forming a link conditional on a meeting taking place. We introduce the notation $p_I \geq p_O$, with the interpretation that, conditional on a meeting, an individual forms the link with probability p_I if the meeting is with someone in his group, and with probability p_O if the meeting is with someone outside his group. Individuals are biased, in the sense that they prefer to connect with similar nodes, when the inequality is strict.[26]

The next result shows that the homophily-degree relationship is flat when the only source of bias is due to individual's preferences, and not to institutional features of the meeting process.

Proposition 12 *Let $p_I > p_O$, and $b_1 = q = 1 - b_2$. Then local homophily $h(\cdot, q)$ is constant.*

Proof. Define the following quantities, where L_{ij} describes the rate of links formed from group i

[26] Currarini, Jackson, and Pin (2009) are independently working on estimating the parameters of a similar notion of choice- versus chance-based biases in the context of the model developed in their 2008 paper.

to group j, and the superscript denotes random- versus network-based search linking processes.

$$\begin{aligned} L_{11}^R &= qp_I b_1 m_r \\ L_{21}^R &= (1-q)p_O(1-b_2)m_r \\ L_{11}^S &= qp_I(\beta_1 b_1 + (1-\beta_2)(1-b_1))m_s \\ L_{21}^S &= (1-q)p_O((1-\beta_2)b_2 + \beta_1(1-b_2))m_r \end{aligned}$$

Then out of all links received by group 1 nodes through the random process, the proportion that come from their own group is $h^R = \frac{L_{11}^R}{L_{11}^R + L_{21}^R}$, and similarly, for links formed through network-based search, it is $h^S = \frac{L_{11}^S}{L_{11}^S + L_{21}^S}$. Let $\alpha = p_I/p_O > 1$. Using $b_1 = q$ and $b_2 = 1-q$, we get that $h^R = \frac{q\alpha}{q\alpha+1-q}$. Next, $\beta_1 b_1 + (1-\beta_2)(1-b_1) = q\beta_1 + (1-q)(1-\beta_2)$ while $(1-\beta_2)b_2 + \beta_1(1-b_2) = q\beta_1 + (1-q)(1-\beta_2)$. As both groups have the same (unbiased) random meetings, they have the same composition of their friends of friends. This implies that $h^S = \frac{q\alpha}{q\alpha+1-q}$. Since $h^R = h^S$, the same level of homophily is achieved in the local neighborhood of an individual independently of which proportion of his links come from one process versus the other. Thus, $h(\cdot, q)$ must be constant, i.e., local homophily is constant in degree. ■

Here random meetings are unbiased, so individuals in both groups meet similar friends of friends. Conversely, the likelihood to be found by a node in the same group is the same for random and search meetings. Even though higher degree nodes still get a higher proportion of their links via search, this has no impact on their local homophily.

Thus, there is a pronounced difference in the structure of networks formed under random meetings bias opposed to under preference biases. One instance of this difference is that the relationship between a node's popularity and the composition of its local neighborhood is qualitatively different in the two cases. Recalling the empirical evidence of Section VI, it is clear that the model with pure-preference biases cannot explain the decreasing relationship between degree and local homophily found with respect to gender in the AddHealth data. Thus, biases in the random meeting process are necessary to capture the main features of that data.

However, there is also reason to believe that preference-based biases are important in link formation. Under pure preference-based biases, for instance, it is straightforward to show that the majority group forms more friends on average than the minority group, a feature which has empirical support (see, e.g., Currarini, Jackson, and Pin (2008)).

Yet another way in which biases might enter the formation process are through the network-based search. It could be that individuals prefer to search in the neighborhoods of similar individuals, or that there are additional institutional constraints that cause them to meet similar nodes among their friends of friends with disproportionate probability. A version of the model without bias in the random meeting process, but instead with biases of the sort described above in the search-based process, would have still different implications for network structure. We conjecture that this model would display an increasing relationship between degree and homophily as, by design, network-based meetings would be more biased than random meetings.

It is clear that different sources of bias in link formation affect the resulting network in empirically testable ways. As much of our focus here is on homophily, we have paid most attention to the effects of these different biases on the relationship between homophily and degree. However, there are likely to be other empirically distinguishing results as well, relating to degree distributions, group-level homophily measures, and other characteristics, such as clustering and network distance.

Our current research aims to develop the predictions of this framework more fully. It is important to map out the effects that different kinds of bias have on other network characteristics. In addition, we are working towards understanding the implications on network structure from the framework that includes all the sources of bias together. There are already some interesting results in this direction. Pin (2009) shows that the local homophily of individuals does not limit to population frequencies in the presence of the additional biases described in this section. Understanding the joint model that accounts for several different sources of bias is useful empirically, as it can describe richer data than the kind we have examined here. For instance, it can accommodate increasing, decreasing, and flat homophily-degree relationships. This can be

useful in interpreting homophily in network data along other dimensions such as race.

IX Concluding remarks

In this paper, we develop a parsimonious model of network formation that accounts for group identity and homophily. This is accomplished by introducing heterogeneity and type-dependent random meeting biases to the model growing random networks proposed by Jackson & Rogers (2007). By using that framework, we start from a model that already accounts for a relatively broad spectrum of network features with strong empirical support. Moreover, it turns out that introducing heterogeneity enriches the analysis significantly. We derive sharp theoretical results regarding the specific patterns of homophily in society and the way homophily interplays with other structural characteristics of the network. In particular, we obtain testable implications on the relationship between individual-level homophily, degree and group composition. We test these implications on data from friendship networks between boys and girls in high schools, finding strong empirical support for the model's predictions. Well-connected students indeed have a more gender-diverse network than those who are relatively isolated.

Our objective is to develop a model that is capable of accommodating rich homophily patterns while keeping the analysis as tractable as possible. One observation that becomes clear is that by introducing a homophilic bias in the very simple way proposed here, the effects on network characteristics can be subtle and complex. However, the analysis has a number of limitations that leave a number of issues open.

First, there are other possible explanations for the empirical relationships that we find. One aspect that falls outside the scope of our model is individual-level heterogeneity. Somewhat specific to the context of adolescent friendships, one could imagine a spectrum of introverted to extroverted individuals. If being extroverted is positively correlated with a dimension of attractiveness, then it would not be surprising that more popular individuals also have a higher proportion of nominations from the opposite gender. While we believe that such individual-level traits are important in determining friendships, the advantage of our approach is that it builds,

in a simple manner, on a framework that captures many other empirically relevant network characteristics. It is not at all clear that starting exclusively from individual-level heterogeneity would allow one to easily recover a natural network structure.

Second, the link formation process that agents follow is specified exogenously; the incentives that might generate such behavior are not modeled. While non-strategic behavior may be a reasonable approximation for studying friendships among adolescents, it may not be appropriate in other contexts. However, this may not be such a large limitation of this modeling approach. As shown by Jackson & Rogers (2007), the decision rule can be viewed as a reduced form outcome of an explicit cost-benefit analysis. Their arguments can be extended to handle the case of group identities studied here. Further, Campbell (2008) presents a dynamic model of strategic interactions that generates the linking decision of Jackson & Rogers (2007), demonstrating explicitly that such a rule can be supported as equilibrium behavior in a fully strategic setting. Finally, while the ability to tie network structure back to micro-level incentives is clearly useful, homophilic biases are observed across many settings, and taking this as fact as a primitive of the model, we generate testable implications that are supported in the data.

Third, our analysis handles only the case of two groups. Thus, we can study homophily with respect to any binary characteristic like gender, but not, by direct application of the model, with respect to characteristics that are continuous (e.g. income) or that have several categories (e.g. race). Many of our results and techniques would extend directly to an analysis of an arbitrary (finite) number of groups, however this comes at the cost of the analysis becoming cumbersome quite quickly.[27]

Finally, in our location-bias model the two groups differ only in terms of their relative proportions in the population. As stated, we intentionally introduced group-dependent linking in its simplest form. However, there are various ways to introduce further group heterogeneity into the model, and these have the potential to significantly enhance the analysis. For instance, the two groups could have different values of γ, m_r or m_s. These effects may be identifiable

[27]For instance, to derive the analog of Proposition 4 with G groups, we have to solve a system of G differential equations in G unknowns for each group. This is not problematic in principle, but is less easy to work with.

empirically as well, depending on the implied characteristics of network structure resulting from such an extension. Augmenting the model to account for such asymmetries could help to explain some of the empirical facts presented in Section VI.

APPENDIX

Consider a degree distribution $F(k)$ obtained implicitly through a process such that $k_i(t) = f(t/i)$ and another degree distribution G such that $k_i(t) = g(t/i)$, with $k_i(i) = 0$. Assume f and g are weakly increasing and continuous on $[1, +\infty[$ and that $\lim_{x\to\infty} f(x) = \lim_{x\to\infty} g(x) = \infty$.

Lemma 1 *F First-Order Stochastically Dominates G if and only if for all $x \geq 1, f(x) \geq g(x)$, with strict inequality for some x.*

Proof. Assume $f(x) \geq g(x)$ for all $x \geq 1$. Pick k and t arbitrarily. Define i_f as the birthdate of the node with degree k at time t under f, and similarly for i_g. We have $f(t/i_f) = k \geq g(t/i_f)$, which, since g is non-decreasing implies that $i_g \leq i_f$. Since $F_t(k) = 1 - i_f/t$ and $G_t(k) = 1 - i_g/t$, we have $F_t(k) \leq G_t(k)$.

Now take $\bar{x}$ such that $f(\bar{x}) > g(\bar{x})$. Pick k and t arbitrarily. Define $i_f = t/\bar{x}$ and $\bar{k}$ to be the size of node i_f at time t under f. Then set i_g to be the node with degree $\bar{k}$ at time t under g. We have $\bar{k} = f(t/i_f) = f(\bar{x}) > g(\bar{x}) = g(t/i_f)$, which implies that $i_g < i_f$. Thus $F_t(\bar{k}) < G_t(\bar{k})$.

To show necessity, fix t and choose i_f so that $f(t/i_f) < g(t/i_f)$, and set $\bar{k} = f(t/i_f)$. Defining i_g as the node with degree $\bar{k}$ at time t under f, we know that $i_g > i_f$. This implies that $G_t(\bar{k}) < F_t(\bar{k})$, completing the proof. ■

Proof. Proposition 4.

Setting $k = \begin{pmatrix} k_{11} \\ k_{12} \end{pmatrix}$, in vector notation, the system of equations to solve is

$$\dot{k} = \frac{1}{t}A + \frac{1}{t}Bk \tag{IX.3}$$

with

$$\begin{aligned} A &= \begin{pmatrix} m\frac{r}{1+r}b_1 \\ m\frac{r}{1+r}(1-b_2)\frac{1-q}{q} \end{pmatrix} \\ B &= \frac{1}{m_r(1+r)} \begin{pmatrix} b_1 m_r & \frac{q}{1-q}(1-b_1)m_r \\ \frac{1-q}{q}(1-b_2)m_r & b_r m_r \end{pmatrix} \equiv \frac{1}{m_r(1+r)} B' \end{aligned}$$

In order to solve equations (IX.3), we transform the system by diagonalizing B. To that end we compute the eigenvalues of B', by writing

$$\det(B' - \lambda I) = \lambda^2 - m_r(b_1+b_2)\lambda + m_r^2(b_1 b_2 - (1-b_1)(1-b_2)),$$

which has solutions

$$\begin{aligned} \lambda' &= \frac{m_r}{2}\left((b_1+b_2) \pm \sqrt{(b_1+b_2)^2 + 4[b_1 b_2 - (1-b_1)(1-b_2)]}\right) \\ &= \frac{m_r}{2}\left(b_1+b_2 \pm (b_1+b_2-2)\right), \end{aligned}$$

which are non-negative if and only if $b_1 + b_2 \geq 1$. Let λ_1 and λ_2 be the eigenvalues of B, i.e., the solutions just computed scaled by $\frac{1}{m_r(1+r)}$, which gives $\lambda_1 = \frac{1}{1+r}$ and $\lambda_2 = \frac{b_1+b_2-1}{1+r}$; set $\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$. Then, there exists a matrix P such that $B = P^{-1}\Lambda P$. Introduce $x = Pk$. Multiplying equation (IX.3) by P leads to

$$\dot{x} = \frac{1}{t} PA + \frac{1}{t}\Lambda x$$

Recall, the solution of $\dot{y} = \frac{a}{t} + \frac{by}{t}$ with $y(i) = 0$ is $y(t) = \frac{a}{b}((\frac{t}{i})^b - 1)$. The initial conditions are

$x(i) = (0,0)'$ since $k(i) = (0,0)'$. This yields

$$\begin{aligned} x_1 &= \frac{(PA)_1}{\lambda_1}[(\frac{t}{i})^{\lambda_1} - 1] \\ x_2 &= \frac{(PA)_2}{\lambda_2}[(\frac{t}{i})^{\lambda_2} - 1], \end{aligned}$$

and since $k = P^{-1}x$,

$$\begin{aligned} k_1 &= \alpha_{11}[(\frac{t}{i})^{\lambda_1} - 1] + \alpha_{12}[(\frac{t}{i})^{\lambda_2} - 1] \\ k_2 &= \alpha_{21}[(\frac{t}{i})^{\lambda_1} - 1] + \alpha_{22}[(\frac{t}{i})^{\lambda_2} - 1] \end{aligned} \tag{IX.4}$$

with $\alpha_{11} = P_{11}^{-1}\frac{(PA)_1}{\lambda_1}$, $\alpha_{12} = P_{12}^{-1}\frac{(PA)_2}{\lambda_2}$, $\alpha_{21} = P_{21}^{-1}\frac{(PA)_1}{\lambda_1}$ and $\alpha_{22} = P_{22}^{-1}\frac{(PA)_2}{\lambda_2}$. We finish the proof by expressing these coefficients as functions of the model's primitives.

In order to compute P^{-1} and P we first need to compute the eigenvectors of B. Set $B'z = \lambda' z$. The first component gives:

$$b_1 m_r z_1 + \frac{q}{1-q}(1-b_1)m_r z_2 = \frac{m_r}{2}\left(b_1 + b_2 \pm (b_1 + b_2 - 2)z_1\right).$$

We can take as a solution $z_1 = -q(1-b_1)m_r$, $z_2^+ = (1-b_2)m_r(1-q)$ and $z_2^- = -(1-b_1)m_r(1-q)$. This yields

$$P^{-1} = \begin{pmatrix} -q(1-b_1)m_r & -q(1-b_1)m_r \\ (1-b_2)m_r(1-q) & -(1-b_1)m_r(1-q) \end{pmatrix}$$

and

$$P = \frac{1}{-q(1-q)(1-b_1)m_r^2(2-b_1-b_2)} \begin{pmatrix} -(1-b_1)m_r(1-q) & +q(1-b_1)m_r \\ -(1-b_2)m_r(1-q) & -q(1-b_1)m_r \end{pmatrix}.$$

Hence

$$\begin{aligned}(PA)_1 &= -\frac{p}{q}\left(\frac{1-b_2}{1-b_1}\right)\frac{1}{2-b_1-b_2}\\ (PA)_2 &= -\frac{p}{q}\left(\frac{b_1+b_2-1}{2-b_1-b_2}\right)\end{aligned}$$

Using the above formulas produces

$$\begin{aligned}\alpha_{11} &= mr\frac{1-b_2}{1-b_1+1-b_2}\\ \alpha_{12} &= mr\frac{1-b_1}{1-b_1+1-b_2}\\ \alpha_{21} &= mr\left(\frac{1-q}{q}\right)\frac{1-b_2}{1-b_1+1-b_2}\\ \alpha_{22} &= -mr\left(\frac{1-q}{q}\right)\frac{1-b_2}{1-b_1+1-b_2}\end{aligned}$$

Substituting these expressions along with the values of λ_1 and λ_2 into equations (IX.4) produces the result. ■

REFERENCES

Ambrus, A. M. Mobius, A. Szeidl (2010) "Consumption Risk-sharing in Social Networks," Working Paper.

Barabási, A. and R. Albert (1999), "Emergence of scaling in random networks," *Science*, **286**: 509-512.

Beaman, L. (2010) "Social Networks and the Dynamics of Labor Market Outcomes: Evidence from Refugees Resettled in the U.S.," Working Paper.

Boschini, A., A. Sjögren (2007), "Is Team Formation Gender Neutral? Evidence from Coauthorship Patterns," *Journal of Labor Economics*, **25(2)**: 325-365.

Bramoullé, Y., R. Kranton (2007) "Risk-Sharing Networks," *Journal of Economic Behavior and Organization*, **64**:275-294.

Calvo-Armengol, T., M. Jackson (2004) "The Effects of Social Networks on Employment and Inequality," *American Economic Review*, **94(3)**:426-454.

Camargo, B., R. Stinebrickner, and T. Stinebrickner (2010), "Interracial Friendships in College," *NBER Working Paper w15970.*

Campbell, A. (2008) "Signaling in Social Networks," mimeo.

Christakis, N., J. Fowler (2008) 'The Collective Dynamics of Smoking in a Large Social Network'," *New England Journal of Medicine*, **358**:2249-2258.

Chung, F., L. Lu (2002a) "The Average Distances in Random Graphs with Given Expected Degrees," *Proceedings of the National Academy of Sciences*, **99**:15879-15882.

Chung, F., L. Lu (2002b) "Connected Components in Random Graphs with Given Degree Sequences," *Annals of Combinatorics*, 6:125-145.

Coleman, J. (1958) "Rational Analysis: The Study of Social Organizations with Survey Methods," *Human Organization*, **17**:28-36.

Conley, T., C. Udry (2010) "Learning About a New Technology: Pineapple in Ghana," *American Economic Review*, forthcoming.

Conley, T., C. Udry (2001) "Social Learning Through Networks: The Adoption of New Agricultural Technologies in Ghana," *American Journal of Agricultural Economics*, **83(3)**:668-673-.

Currarini, S., M. Jackson, P. Pin (2008) "An Economic Model of Friendship: Homophily, Minorities and Segregation," *Econometrica*, forthcoming.

Currarini, S., M. Jackson, P. Pin (2009) "Identifying the Roles of Choice and Chance in Network Formation: Racial Biases in High School Friendships," mimeo.

Fafchamps, M., S¿ Lund (2003) "Risk Sharing Networks in Rural Philippines," *Journal of Development Economics*, **71**:261-287.

Golub, B., M. Jackson (2008) "How Homophily affects Communication in Networks," mimeo.

Jackson, M. (2008a) "Average Distance, Diameter, and Clustering in Social Networks with Homophily," to appear in the *Proceedings of the Workshop in Internet and Network Economics*

(WINE 2008), Lecture Notes in Computer Science edited by C. Papadimitriou and S. Zhang, Springer-Verlag, Berlin Heidelberg.

Jackson, M. (2008b) "Social and Economic Networks," Princeton University Press.

Jackson, M., B. Rogers (2007) "Meeting Strangers and Friends of Friends: How Random are Social Networks?," *The American Economic Review*, 97(3).

Laumann, E., Y. Youm (1999) "Racial/ethnic group differences in the prevalence of sexually transmitted diseases in the United States: a network explanation," *Sexually Transmitted Diseases*, **26(5)**:250-261.

Lazarsfeld, P., and R. K. Merton (1954) "Friendship as a Social Process: A Substantive and Methodological Analysis," In Freedom and Control in Modern Society, Morroe Berger, Theodore Abel, and Charles H. Page, eds. New York: Van Nostrand, 18-66.

Mayer, A., and S. Puller (2008) "The Old Boy (and Girl) Network: Social Network Formation on University Campuses," *Journal of Public Economics,* **92**(1-2), 329-347.

Montgomery, J. (1991) "Social Networks and Labor-Market Outcomes: Toward an Economic Analysis," *American Economic Review*, **81**:1408-1418.

Newman, M. (2003) "The structure and function of complex networks," *SIAM Review*,**45**, 167-256.

Newman, M. (2004) "Coauthorship networks and patterns of scientific collaboration," *Proceedings of the National Academy of Sciences*, **101**: 5200-5205.

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

Pin, P. (2009) "Four multi-agents economic models: From evolutionary competition to social interaction," PhD Thesis, VDM Verlag.

Vigier, A. (2008) "Network Topology: Extracting Economic Content," *mimeo.*

Watts, D. and S. Strogatz (1998) "Collective dynamics of 'small-world' networks," *Nature,* **393**: 440-442.