Diversity and Popularity in Social Networks

Yann Bramoullé and Brian W. Rogers*

January 19, 2009

Abstract: Homophily, the tendency of linked agents to have similar characteristics, is an important feature of social networks. We present a new model of network formation that allows the linking process to depend on individuals types and study the impact of such a bias on the network structure. Our main results fall into three categories: (i) we compare the distributions of intra- and inter-group links in terms of stochastic dominance, (ii) we show how, at the group level, homophily depends on the groups size and the details of the formation process, and (iii) we understand precisely the determinants of local homophily at the individual level. Especially, we find that popular individuals have more diverse networks. Our results are supported empirically in the AddHealth data looking at networks of social connections between boys and girls.

JEL Codes: A14, D85, I21.

*Bramoullé: Department of Economics, CIRPÉE and GREEN, Université Laval.
Rogers: MEDS, Kellogg School of Management, Northwestern University.
We thank Andrea Galeotti, Sanjeev Goyal, Matthew Jackson, Bruno Strulovici, and Adrien Vigier, as well as seminar participants at the First Transatlantic Theory Workshop (Paris), The University of Essex, and Cambridge University for helpful comments and conversations. 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

A thorough understanding of homophily is essential to the study of social and economic networks. Homophily is the tendency for similar individuals to be connected to each other. While its prevalence has been documented empirically across many contexts, it has been relatively neglected, so far, in theoretical studies on networks. Homophily implies positive correlations in the characteristics of individuals linked to each other in a network, and such correlations may have substantial implications for the operation of various processes within the network. The outcomes regarding diffusion of information, technologies, and disease, behavioral outcomes in games of strategic complements or substitutes played in a networked society, and observational learning may all depend non-trivially on homophilic aspects of the network.

We develop a model of network formation in order to advance our understanding of homophily. The framework allows us to keep track of the characteristics of individuals, measured as a group identity, and how they relate to their pattern of connections in the network. We are able to uncover some systematic relationships that are not obvious a priori. We are especially interested in the relationship between degree (number of connections), degree distributions across society, and the extent of local and aggregate homophily. To investigate these relationships, we extend the model of Jackson & Rogers (2007). This model provides an ideal starting point as it is both tractable and able to replicate most empirically documented features of social networks *except for homophily*. This paper overcomes this limitation, yielding new insights and empirical predictions.

Notably, we show that *individuals with a higher degree should have more balanced neighborhoods*. This prediction turns out to be strongly supported empirically, when looking at gender bias in friendship networks among adolescents. Using Add Health data, we find that in high school, boys and girls who receive more friendship nominations indeed have a much more diverse network of friends.

This paper contributes to a growing literature studying stochastic models of network formation. Until very recently, most studies have focused mainly on degree distributions and other summary statistics of the network independent of individuals' types or other characteristics, see

for instance Barabasi and Albert (1999), Chung & Lu (2002a,b), Jackson & Rogers (2007), Newman (2003, 2004), and Watts and Strogatz (1998). Building on this literature, we acknowledge the empirical importance of community structure and investigate its interplay with the other network features.

Two other recent papers have started to tackle these issues.[1] Currarini, Jackson, and Pin (2008) 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 quite different from ours. Interestingly, however, a number of their predictions are also supported by our model. Jackson (2008) incorporates homophily into the random graph model of Chung & Lu (2002a,b).[2] Again by design, homophily cannot vary with degree in this approach. Also, degree distributions constitute an outcome of our model while they are an input of Jackson (2008). Thus, our analysis 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 individual's degree.

We now describe our model in more detail. As in Jackson & Rogers (2007), individuals are born sequentially. Upon entry, an individual meets others in two distinct ways: at random and through network-based search. The network-based search allows individuals to meet the neighbors of their randomly met partners. In either case, conditional on a meeting taking place, a connection is formed with some probability. We depart from Jackson & Rogers (2007) by assigning individuals a group identity or label, and allowing the meeting process to be affected by group identity.[3] Specifically, we assume that random meetings take place in either of two "locations" and that nodes of a specific group are exogenously more likely to be in a specific location. These locations could indeed be spatially based, such as geographic neighborhoods, or

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

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

[3] A different approach by Vigier (2008) extends the Jackson & Rogers (2007) model to allow link formation to depend on an exogenous metric that allows for homophily.

instead they could represent more general social interaction settings, including church and club memberships. We leave the rest of the network formation process unchanged and study how this bias in random meetings propagates through the system.[4] This allows us to develop a clean understanding of homophily patterns induced by random meeting biases.[5]

We obtain three sets of theoretical results about homophily, degree distributions and the relation between homophily and degree, which we now outline.

First, we study how the exogenously taken bias in random meetings relates to the observed homophily in society as the population grows large. Our main finding here is that network search has a dampening effect on homophily. Meeting "friends of friends" opens up more possibilities for diverse relationships. Thus, network homophily decreases as the ratio of search-based to random meetings increases. We also look at the relationship between the relative homophily within a particular group and the relative size of that group in society. As in Currarini, Jackson, and Pin (2008), we find that this relation exhibits an inverted U-shape. It is remarkable that two models built on very different premises lead to this similar conclusion.

Second, we use the mean-field approximation to characterize the degree distributions of the social network. In the presence of community structure, we are interested in seven distributions rather than only one. For instance, we can keep track of the distribution of the number of links that boys have with girls, boys have with boys, and so forth. We study the shapes and comparative statics of these distributions. We show that they all have a scale-free upper tail *with a common exponent*, even though the distributions themselves are not scale-free. We also derive a number of comparisons across distributions using first-order stochastic dominance. Finally, the distribution of total in-degree for two populations are identical, independent of the relative group sizes and extent of location biases.

Third, we examine the relationship between an individual's degree and the homophily in its immediate neighborhood. That is, we examine how the number of links of a node is related to the proportion of its neighbors that belong to the same group. Our main result here is that

[4]We do not introduce a group-dependent bias in link formation conditional on meeting. We return to discussing this issue in the conclusion.

[5]In fact, we derive many of our results under a general representation of random meeting biases that encompasses the location-based meetings described here.

this local homophily decrease with an individual's degree. In other words, nodes with a higher degree have more diverse networks. In the limit, the homophilic bias completely disappears as degree becomes very high, independent of the bias in random meetings. In addition, we find that this decrease takes place at a decreasing rate and is less pronounced in larger groups. This analysis provides the first comprehensive results relating homophily and degree distributions in social networks.

Our final contribution is to bring data to bear on these novel predictions. We look at gender and friendship networks among high school students in the U.S. provided in the AddHealth data. First, we find that the distribution of in-degree is remarkably similar for boys and girls. Second, we test the hypotheses that individual homophily is decreasing and convex in degree, and that the relationship is more dramatic in smaller groups. Regressions yield highly significant coefficients for our variables of interest, all with the predicted signs. These findings may have important consequences for the debate on gender composition and social interactions at school.

The paper proceeds as follows. The next section describes the formal model. Section III contains the fundamental results on homophily. Section IV derives degree distributions for the network under the mean-field approximation, and presents a number of relationships across degree distributions. Section V describes how local homophily varies with degree, which highlights the importance of studying a model with non-trivial 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. The final section offers concluding remarks. An appendix contains some supplemental formal results.

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 $\gamma \in [.5, 1]$, which represents the probability that a g^i node goes to location L^i, $i = 1, 2$.[6] Thus, it is simply the resulting composition of types in the two locations that permits any group-dependent biases in the model. Once at a particular location, all 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

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

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 are able to prove many of our results in a more general setting in which b_1 and b_2, taken as primitives of the model rather than implications of location-based biases, 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* 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]

[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.

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 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 does not rely on assuming that the homophilic biases are driven by location choices, and is true for any specification 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 some elaboration. Recall that through search a node has access to all the parents' neighbors. This includes, of course, the neighbors of parents in the other group which, because of homophily, tend to also belong to the other group. Thus, search gives access to many nodes of the other group. In short, meeting the friends of distinct friends opens up possibilities for diverse relationships. Search acts here as a countervailing force with regards to the original 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]

Next, we want to 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. The first 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 $H_1(0) = H_1(1) = 0$. The first derivative of H_1 is

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

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

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

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

[8] We show below that this parameter also plays a crucial role in degree distributions.

[9] Conversely, we can show that the model always exhibit network inbreeding homophily only when $\beta_{10} = q$ and $\beta_{20} = 1 - q$.

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

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. 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, and 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 search-based meetings are 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 Relationships among the distributions of links

In this section, we make use of the powerful mean-field approximation to study degree distributions. Even with only two groups, the distribution of links becomes substantially more complex than the Jackson & Rogers (2007) case as nodes can connect to both same and different type nodes, and one wants to keep track of the different kinds of links. We focus on the distributions of in-degrees, as out-degrees are essentially homogenous, with all nodes of the same type forming their out-links according to the same (stochastic) rule. 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 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 8 that we specialize to consider biases driven by γ and location choices.

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.

Now consider the individual entering at time $t > i$. What are the probabilities that this node forms a link with i, and hence that k_{11}^t or k_{12}^t increase by one unit?

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

The new node 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 $\frac{m_s}{mm_r}$, 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

$$P(k_{12}^{t+1} = k_{12}^t + 1) = (1-q)p(\frac{(1-b_2)m_r}{qt} + \frac{m_s}{mm_r}(k_{11}^t\frac{(1-b_2)m_r}{qt} + k_{12}^t\frac{b_2m_r}{(1-q)t})).$$

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. We have

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).[11]

To obtain F_{11}, observe that $F_{11}(k) = 1 - i/t$, which means that $t/i = 1/(1 - F_{11}(k))$. This defines F_{11} 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 5 *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, and in particular, that they tend to a scale-free distribution with exponent $-(1 + r)$.

Proposition 6 *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 $F_{11}(k)$:

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

[11] Both 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.

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 fat 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 empirical 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 7

(i) When $q > 1/2$, F_{11} FOSD F_{12}.

(ii) If $q > 1/2$ and $b_1 + b_2 < 2$, 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 becomes

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

■

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 neighbors. In other words, the hubs in the minority group have the greatest proportion of their in-neighbors from the majority group. This effect is proven in Section V.

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 8 *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 7 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 next result describes how the distributions of inter- and intra-group links respond to changes in the homophilic bias.

Proposition 9 *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.

V The relationship between local homophily and degree

In this section, we want to analyze how the local homophily surrounding an individual varies with its degree. Since out-degree is constant, we look at in-degrees.

We define *node i homophily* as the proportion of i's in-neighbors who belong to i's group. When i belongs to group g^1, this is equal to $k_{11}(t)/k_1(t)$. We show first that individual homophily decreases with degree, which again applies without the restriction that random meetings are location-driven. We then obtain sharper results under location-driven meetings.

Proposition 10 *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 $h^j(k)$. Then, $\forall k > 0, (h^j)'(k) < 0$ 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 turns out to be independent of t and i. Next, compute the derivative of h: $h'(k) = (f^{-1})'(k)g'(f^{-1}(k))$. Here, $(f^{-1})' > 0$ and $g'(x)$ 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 are always relatively less homophilic in their immediate neighborhoods. The intuition for the result comes, again, from the differences between the two ways to form links. 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. 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 even cancels out any bias originally coming from the random meetings.

Next, we look at location-based random meetings. This allows us to derive three additional properties on the relation between degree and homophily.

Proposition 11 *Suppose that biases are location-driven. Then,*

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

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

Proof. Here, $f(x) = mr(x^{\alpha_1} - 1)$ 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 biases are location-driven. Homophily is also higher in larger groups. Recall the discussion after Proposition 7 focusing on the interplay between homophily and the size effect of a group. Members of a relatively larger group, all else equal, will have a higher proportion of intra-group links due to the size effect. Perhaps more surprisingly, the relation between homophily and degree is also *less* decreasing in larger groups. Alternatively, 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

Before proceeding, we remark that our theoretical analysis relies on a mean-field approximation. Therefore, we conducted a set of computer simulations of the underlying network formation process. This allows us to check directly the validity of the predictions against the results of the simulations.

Each simulation was run under location-driven biases for $T = 3000$ nodes with $m_r = 10$. We varied b_1, b_2, m_r, m_s, p, q. This allows us to have enough variation in the key parameters to have some confidence in the global predictions of the model. Notice that this generates variation in γ as well, since its value is pinned down by b_1, b_2, and q when biases are location-driven.[12]

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

[12]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 one. 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 $q \geq \frac{1}{2}$.

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 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, these network constraints do not 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 up to $10,000$ 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 V are most central to our motivating questions, we focus on discussing the relationship between degree and individual homophily in the simulated networks.[13] In summary, we find strong support for the model's predictions. Look at Figure 1, which depicts the relationship between individual homophily and degree for a typical simulation. Each point represents the average individual homophily among nodes of a particular in-degree. The bounds from Proposition 11 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.

VI 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

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

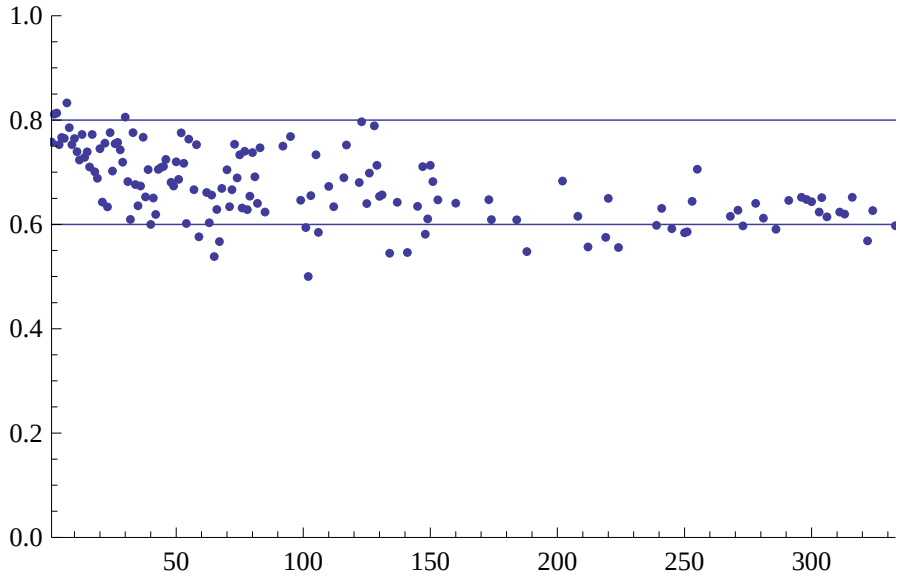

Figure 1: *Individual homophily as a function of in-degree for a typical simulation. The parameters here are* $q = 0.6$*,* $b_1 = .8$*,* $b_2 = .7$*,* $r = \frac{1}{3}$ *and* $p = 1$*. The implied value of* γ *is approximately* 0.86*.*

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 10 and 11 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. 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 him 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 him 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.[14] Very few links originating 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 variations. First, we have some variation across schools in terms of gender composition. Second, we have natural variation in degrees of students within schools. Figure 2 depicts the frequency distribution of in-degrees in the data.

We now look at the relationship between homophily and degree at the individual level. Figure 3 depicts how average individual homophily varies with in-degree over the whole sample.[15] The relation clearly shows a decreasing and convex shape. The effect of degree also seems to be 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.[16] Our analysis in section V shows that under the Location-Bias 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.

[14]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.

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

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

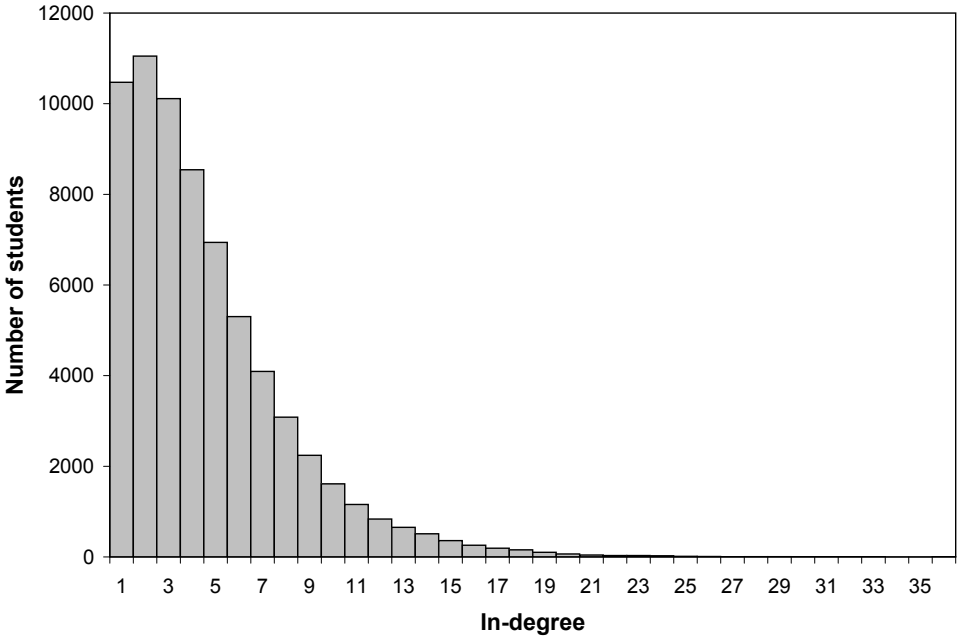

Figure 2: *The distribution of in-degree across students in the AddHealth sample.*

	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
	$(4 * 10^{-5})$	$(6 * 10^{-5})$	$(6 * 10^{-5})$
q^j	0.53	0.567	0.268
	(0.018)	(0.017)	(0.087)
degree$*q^j$	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

Table 1. *Regression estimates of individual homophily on degree, degree^2, q, and an interaction term degree$*q$. Robust standard errors are given in parentheses.*

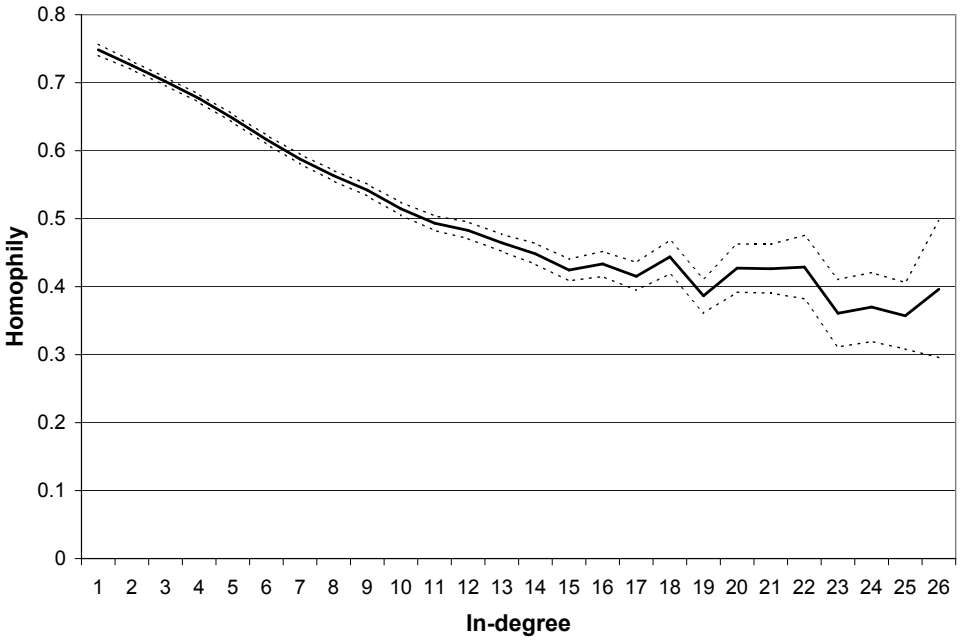

Figure 3: *The relationship between individual homophily and in-degree shows a decreasing and convex pattern.*

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.[17] 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.

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

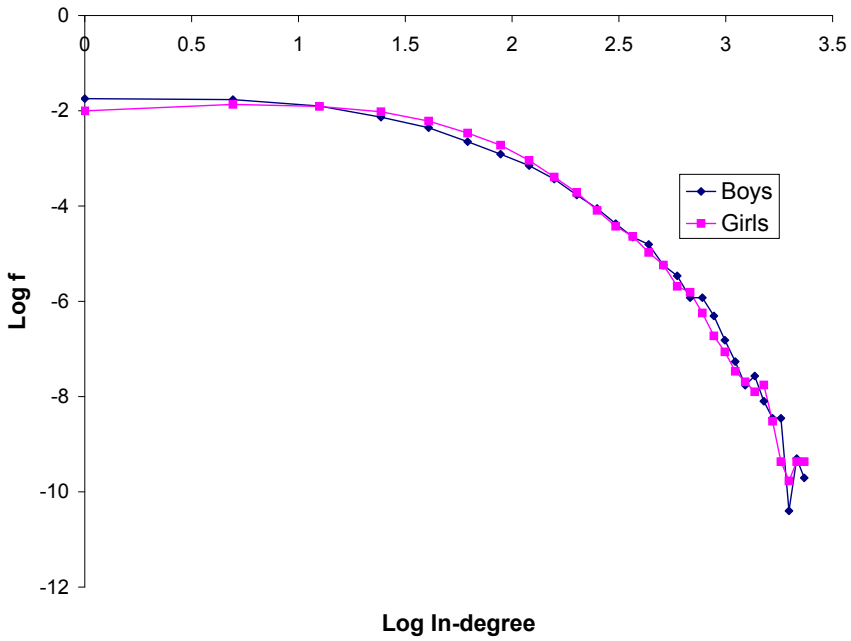

Figure 4: *The pdf of in-degrees for boys (blue) and girls (pink) shown in log-log scale.*

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. 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.[18] In summary, we find strong empirical support for our theoretical predictions relating individual homophily to other network features.

The model yields further testable implications. In section IV, we derive several results constraining the relationships among various degree distributions. However, the AddHealth data

[18]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.

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 8). 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 4 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.[19]

Finally, in section III, we show that the relationship between relative homophily and school gender composition should have an inverted-U shape.[20] To check for this relationship, we regress H_j on q_j and q_j^2 over all schools. This yields coefficients that are statistically not significant.[21] 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.[22] A flat relationship between homophily and gender composition around $q = 0.5$ is consistent with Proposition 8. However, the range is probably too small to lead to an informative test of the result.

VII Conclusion

In this paper, we develop a parsimonious model of network formation that accounts for group identity and homophily. This is accomplished by introducing type-dependent random meeting biases to the model of Jackson & Rogers (2007). It turns out that doing so enriches the analysis significantly. In particular, we derive sharp theoretical results regarding the specific patterns of homophily in society and the way homophily interplays with other structural characteris-

[19]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 18, provides an interesting direction for future research; see the conclusion.

[20]Currarini, Jackson and Pin (2008) find empirical evidence of this inverted-U shape on the same data, but with respect to racial homophily.

[21]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.

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

tics 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 open questions for future research.

First, 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 reasonable for studying friendships among adolescents, it may not be appropriate in other contexts. We have three comments on this point. First, 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. Second, Campbell (2008) presents a dynamic model of strategic interactions that generates the linking decision of Jackson & Rogers (2007), demonstrating that such a rule can be supported as equilibrium behavior in a fully strategic setting. Third, 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.

Second, 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.[23]

Third, 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. In principle these effects may be identifiable empirically as well. Extending the model to account for such asymmetries could help to explain some of the empirical facts presented in Section VI.

Finally, we abstract away from potential biases in the probabilities of initiating a link conditional on meeting. In some settings, observed homophily may arise from a bias in interaction opportunities, as in our model, or instead from a bias in preferences. That is, one could consider a model in which the probability of connecting to an agent depends on whether or not the two individuals belong to the same group, thereby introducing a second group-dependent bias. Our model allows us to provide a clean analysis of how homophily and network formation are affected by biases in the meeting process only. Examining the data, the analysis shows that empirical patterns of homophily may be well explained by such opportunity bias. Still, both types of biases may matter in other contexts. Empirically identifiying the precise source of homophily is an important question. A promising direction for future work is to account for both types of biases in a common framework, to study their interaction and their joint empirical implications.

[23]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.

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

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

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 (VII.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 (VII.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 $k = P^{-1}x$,

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

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 (VII.4) produces the result. ■

REFERENCES

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

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

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.

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

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

Jackson, M. (2008) "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., B. Rogers (2007) "Meeting Strangers and Friends of Friends: How Random are Social Networks?," *The American Economic Review*, 97(3).

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.

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.

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.