

Does Homophily Predict Consensus Times? Testing a Model of Network Structure via a Dynamic Process*

Benjamin Golub[†]

Matthew O. Jackson[‡]

Draft: July 27, 2012

Forthcoming in the *Review of Network Economics*

Abstract

We test theoretical results from Golub and Jackson (2012a), which are based on a random network model, regarding time to convergence of a learning/behavior-updating process. In particular, we see how well those theoretical results match the process when it is simulated on empirically observed high school friendship networks. This tests whether a parsimonious random network model mimics real-world networks with regard to predicting properties of a class of behavioral processes. It also tests whether our theoretical predictions on asymptotically large societies are accurate when applied to populations ranging from thirty to three thousand individuals. We find that the theoretical results account for more than half of the variation in convergence times on the real networks. We conclude that a simple multi-type random network model with types defined by simple observable attributes (age, sex, race) captures aspects of real networks that are relevant for a class of iterated updating processes.

Keywords: homophily, friendships, social networks, random graphs, random network models, spectral graph theory, social learning

JEL Classification Numbers: C51, D85

*We are grateful to Andrea Galeotti and an anonymous referee for very helpful suggestions. Jackson gratefully acknowledges financial support from the NSF under grants SES-0961481 and SES-1155302. Golub gratefully acknowledges financial support from an NSF Graduate Research Fellowship. This article includes material that originally appeared in the working paper “How Homophily Affects Diffusion and Learning in Networks” (arXiv:0811.4013v2).

[†]Department of Economics, MIT. Email: bgolub@mit.edu, web: <http://web.mit.edu/bgolub/www/>.

[‡]Department of Economics, Stanford University and external faculty of the Santa Fe Institute. Email: jacksonm@stanford.edu, web: <http://www.stanford.edu/~jacksonm/>

1 Introduction

The structure of social networks affects the functioning of a society and an economy in essential ways, especially when it comes to the flow of information and the dynamics of behaviors. Models of processes and games that take place on a network relate network structure to outcomes. Given a complete description of the network, such models can make precise, testable predictions about quantities of interest. For example, DeGroot’s (1974) model of repeated linear updating of beliefs or behaviors fully specifies the dynamics of outcomes given a network structure and initial conditions.¹

However, networks are inherently complicated objects, and so from both a conceptual and empirical perspective it is desirable to have results in terms of statistics that are lower-dimensional, more interpretable, and easier to estimate than the full structure of a network. Therefore, researchers often model a network as some sort of random graph described by a relatively small number of parameters – ideally, ones that are substantively meaningful and can be estimated. Then the behavior of interest can be related to these parameters rather than the full network structure, potentially yielding more insight and more empirical traction.

This raises an important question about the extent to which a model of random networks proxies for an actual network when a behavior happening on the network is analyzed. In particular, we know that any model of networks must miss some characteristics of a real network, and thus we need to check that if we calculate the dynamics of a process of interest on the modeled network, they are similar to simulations of the process on a real-world network. In this paper we investigate that question in the context of a specific but important behavioral process and class of random networks.

In particular, we consider a *multi-type random network* model, a variation of standard stochastic block models, that was used by Golub and Jackson (2012a) as a foundation upon which to study a learning and updating process. In that random

¹The essence of the DeGroot model is that nodes in a network update their states by taking averages of the states of their neighbors. This model was first discussed by French (1956) and Harary (1959), and later presented in its general form by DeGroot (1974). Examples of analysis of the dynamics include DeMarzo, Vayanos, and Zwiebel (2003); Golub and Jackson (2010); and Jadbabaie and Tahbaz-Salehi (2010) within a literature in engineering and control. Some recent work, including Jadbabaie et. al (2012) and Mueller-Frank (2012), studies how the dynamics can differ when updating rules are more intricate than pure linear updating but still have an averaging aspect. Empirical investigations of such updating processes are offered by Chandrasekhar, Larreguy, and Xandri (2010) and by Corazzini et al. (2011). Some discussion of other related literature is provided in Jackson (2008).

network model, a society consists of an arbitrary number of groups, or types, each making up an arbitrary fraction of society. A link forms between two agents with a probability that is determined by the exogenously defined types (e.g., demographic characteristics) of both of the agents involved; links are conditionally independent given these type-based probabilities. In this setting, Golub and Jackson (2012a) examined the DeGroot updating process and specifically the persistence of disagreement (formally defined by a measure we call *consensus time*). We showed that consensus time is characterized in terms of a certain measure of segregation, called *spectral homophily*, that depends only on large-scale linking patterns among groups, and (with high probability) not on idiosyncratic details of network realizations. Essentially, the spectral homophily measures the extent to which agents of a given type are biased toward forming links with other, similar types.² This quantity can be estimated by taking a small sample of agents and considering the relative frequencies of links: for example, how many Asian friends an Asian high school student tends to have for every white friend.

The usefulness of such results depends on whether the DeGroot process operating on actual networks acts similarly to the DeGroot process on the multi-type random graph model. In addition, the results are of practical interest only if agent “types” that determine linking probabilities are qualities that can be observed by researchers. More generally, whenever a potentially complicated characterization of behavior in a network is reduced to a simple invariant that can be estimated using a few large-scale parameters, there is a question of whether the assumptions underlying that reduction are valid for the data in question.

With this in mind, we examine social networks from over eighty different high school friendship networks from the National Longitudinal Study of Adolescent Health (Add Health³). Given full network data, we are able to simulate a DeGroot updating process on the actual networks and calculate worst-case consensus times. In addition to full network data, we have the grade in school, sex, and race of each student in each network. The networks exhibit variation in numbers of nodes, density of links, and homophily. As a result, we are able to check whether the predictions of our

²See also Golub and Jackson (2012b) for discussion of a related measure of homophily defined in terms of the structure of the network.

³Add Health is 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, with cooperative funding from 17 other agencies. 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).

theorems hold – that is, whether consensus time depends on a global measure of homophily from the multi-type random graph model. We should expect this to be the case only if the multi-type random graph model – with the types defined by the demographic attributes available in the data – captures essentials of the structure of actual networks for the purpose of an analysis of (DeGroot) belief updating.

Since we do not employ data on the actual process of communication, our exercise does *not* test whether learning in reality follows an average-based process⁴. However, it does test whether the multi-type random network captures some aspect of social network structure, since otherwise there is little reason to expect the theoretical predictions to hold even on simulated communication processes.

We find that the speeds of linear updating processes depend on homophily in ways that are consistent with the theory. Indeed, variation in spectral homophily based on simple observed attributes predicts more than half of the variation in consensus times observed in the data. In view of this, we conclude that the multi-type random graph model based on grade-sex-race types captures at least some of the important aspects of the real social network structures, at least for purposes of studying linear updating processes. From a substantive perspective, a simple calculation of consensus time based on group-level linking tendencies can, in some settings, be a good predictor of the communication dynamics arising from potentially intricate micro-level patterns.

2 The Model

We work with a simple model of network structure that allows for arbitrary large-scale patterns of interaction among groups, while making individual (stochastic) link realizations independent conditional on the types of agents involved. We provide an abbreviated description that is based on Golub and Jackson (2012a), and refer the reader to that paper for more detail.

2.1 Multi-Type Random Networks

Given a set of n nodes $N = \{1, \dots, n\}$, a network is represented via its adjacency matrix: a symmetric n -by- n matrix \mathbf{A} with entries in $\{0, 1\}$. The interpretation is that $A_{ij} = A_{ji} = 1$ indicates that nodes i and j are linked, and the symmetry restricts

⁴See Chandrasekhar, Larreguy, and Xandri (2010) and Corazzini et al. (2011) for analyses in that direction.

attention to undirected networks.

Let $d_i(\mathbf{A}) = \sum_{j=1}^n A_{ij}$ denote the degree (number of links) of node i , the basic measure of how connected a node is. Let $d(\mathbf{A})$ denote average degree. Finally, let

$$D(\mathbf{A}) = \sum_i d_i(\mathbf{A})$$

be the sum of all degrees in the network, which is twice the total number of links.

Agents or nodes have “types,” which are the distinguishing features that affect their propensities to connect to each other. Types might be based on any characteristics that influence agents’ probabilities of linking to each other, including age, race, gender, profession, education level, and even behaviors.⁵ For instance, a type might consist of the eighteen year-old female African-Americans who have completed high school, live in a particular neighborhood, and do not smoke. The model is quite general in that a type can embody arbitrary lists of characteristics; which characteristics are included will depend on the application. There are m different types in the society. Let $N_k \subset N$ denote the nodes of type k , so the society is partitioned into the m groups, (N_1, \dots, N_m) . Let $n_k = |N_k|$ denote the size of group k ; denote by $\mathbf{n} = (n_1, \dots, n_m)$ the corresponding vector of cardinalities; and let n denote the total number of agents.

A *multi-type random network* is defined by the cardinality vector \mathbf{n} together with a symmetric m -by- m matrix \mathbf{P} , whose entries (in $[0, 1]$) describe the probabilities of links between various types.⁶ The resulting random network is captured via its adjacency matrix, which is denoted by $\mathbf{A}(\mathbf{P}, \mathbf{n})$.⁷ In particular, $\mathbf{A}(\mathbf{P}, \mathbf{n})$ is built by letting the entries A_{ij} with $i > j$ be independent Bernoulli random variables that take on a value of 1 with probability $P_{k\ell}$ if $i \in N_k$ and $j \in N_\ell$. That is, the entry $P_{k\ell}$ captures the probability that an agent of type k links to an agent of type ℓ . We fill in the remaining entries of $\mathbf{A}(\mathbf{P}, \mathbf{n})$ by symmetry: $A_{ij} = A_{ji}$. Unless otherwise noted, we use $\mathbf{A}(\mathbf{P}, \mathbf{n})$ to denote a random matrix, and \mathbf{A} without an argument to refer to a given deterministic matrix.⁸

⁵However, we do not allow types to depend on behaviors or beliefs that are endogenous to the updating being modeled.

⁶We assume a numbering of agents such that $N_1 = \{1, 2, \dots, n_1\}$, $N_2 = \{n_1 + 1, n_1 + 2, \dots, n_1 + n_2\}$, and so on. Given this convention, it is possible to recover the partition from just the vector \mathbf{n} .

⁷This kind of stochastic network with type-dependent heterogeneity can arise from a strategic friendship formation process with some randomness in the order of meetings, as in Currarini, Jackson, and Pin (2009).

⁸For individual entries, we drop the arguments (\mathbf{P}, \mathbf{n}) , but the matrix in question will be clear

The multi-type random network model is a standard stochastic block model (e.g., see Jackson, 2008, for references) and subsumes many other random network models. The seminal random network model of Erdős and Rényi is a special case, as are many cases of the model based on degree distributions of Chung and Lu (2002).⁹ One can also view the probabilities in the matrix \mathbf{P} as arising from distances between some underlying locations, either physical or abstract. While the model as presented restricts locations to be discrete, it would be feasible to permit a continuous specification. One can even include different sociabilities, so that some groups consist of individuals who, for whatever reasons, form more relationships on average than others. Thus, it need not be that all nodes have the same expected number of connections; the network can have a nontrivial degree distribution.¹⁰

2.2 A General Measure of Homophily

We now provide a general definition of homophily based on the probabilities of interaction between various types, and then show how it works in an important special case.

Let $Q_{k\ell}(\mathbf{P}, \mathbf{n}) = n_k n_\ell P_{k\ell}$ be the expected total contribution to the degrees of agents of type k from links with agents of type ℓ ; when $k \neq \ell$, this is simply the expected number of links between k and ℓ . Also, let $d_k(\mathbf{Q}(\mathbf{P}, \mathbf{n})) = \sum_\ell Q_{k\ell}(\mathbf{P}, \mathbf{n})$ be the expected total degree of nodes of type k .

Let $\mathbf{F}(\mathbf{P}, \mathbf{n})$ be a matrix of the same dimensions as \mathbf{P} with entries

$$F_{k\ell}(\mathbf{P}, \mathbf{n}) = \frac{Q_{k\ell}(\mathbf{P}, \mathbf{n})}{d_k(\mathbf{Q}(\mathbf{P}, \mathbf{n}))}. \quad (1)$$

(We take $0/0 = 0$ throughout the paper.) Thus, $F_{k\ell}$ is the expected fraction of their links that nodes of type k will have with nodes of type ℓ . This simplifies things in

from context.

⁹This can also be seen as a cousin of some of the statistical models that have been used to capture homophily in networks, such as some \mathbf{p}^* and exponential random graph models (e.g., see the references and discussion in Jackson, 2008). There are variations on it in the computer science literature called the planted multi-section model, e.g., McSherry (2001). An early version of this type of model was introduced by Diaconis and Freedman (1981) in a study on the psychology of vision, independently of its introduction in the stochastic block modeling literature (e.g., Fienberg and Wasserman, 1981, and Holland, Laskey, and Leinhardt, 1983), which provided a basis for some of the community detection literature, e.g., Copic, Jackson, and Kirman (2009).

¹⁰There are, of course, networks that this approach is not well suited to modeling: strict hierarchies, perfect regular lattices, etc., even though homophily can and does occur in these networks.

two respects relative to the realized random network. First, it works with groups (or representative agents of each type) rather than individual nodes; and second, it works with ratios of expected numbers of links rather than realized numbers of links. With this matrix defined, we can formulate a general homophily measure.

DEFINITION 1. The *spectral homophily* of a multi-type random network (\mathbf{P}, \mathbf{n}) is the second-largest eigenvalue¹¹ of $\mathbf{F}(\mathbf{P}, \mathbf{n})$. We denote it by $h^{\text{spec}}(\mathbf{P}, \mathbf{n})$.

The spectral homophily measure is based on: (i) simplifying the overall interaction matrix to that of the expected interaction across groups, and (ii) then looking at a particular part of the spectrum of that matrix: the second-largest eigenvalue. On an intuitive level, a second-largest eigenvalue captures the extent to which a matrix can be broken into two blocks with relatively little interaction across the blocks. Indeed, in Section VI of Golub and Jackson (2012a), we present a formal result showing that spectral homophily picks up “fault lines” created by segregation in the network. In the next section, we illustrate this in the context of a special case.

2.3 A Special Case: The Islands Model

For an illustration of the general definitions, it is useful to consider a situation in which groups are equal-sized and they are all biased in the same way. In particular, links within a type are more probable than links across types, and the probability of a link across types does not depend on the specifics of the types in question. This is a case that we call the “islands” model.

More precisely, the *islands model* is the special case of the multi-type random networks model such that (i) each type (group) has the same number of agents; and (ii) an agent only distinguishes between agents of one’s own type and agents of a different type. Moreover, all agents are symmetric in how they do this. Formally, in the multi-type random network notation, we say the multi-type random network (\mathbf{P}, \mathbf{n}) is an islands network with parameters (m, p_s, p_d) if:

- there are m groups (islands) and their sizes, n_k , are equal for all k ;

¹¹As shown in Golub and Jackson (2012a), all eigenvalues of $\mathbf{F}(\mathbf{P}, \mathbf{n})$ are real. To define the second-largest eigenvalue, list the eigenvalues of this matrix ordered from greatest to least by absolute value, with positive eigenvalues listed first if there are ties. An eigenvalue is listed the same number of times as the number of linearly independent eigenvectors it has. The second eigenvalue in the list is called the second-largest eigenvalue. (Clearly, its magnitude does not depend on the tie-breaking assumption.)

- $P_{kk} = p_s$ for all k ; and
- $P_{k\ell} = p_d$ for all $k \neq \ell$, where $p_d \leq p_s$ and $p_s > 0$.

The idea that agents only distinguish between “same” and “different” agents in terms of the linking probabilities is surprisingly accurate as a description of some friendship patterns (e.g., see Marsden, 1987, and footnote 7 in McPherson, Smith-Lovin, and Cook, 2001).

Figure 1 depicts two different random networks generated by the islands model, with different linking probabilities.

In the context of the islands model, it is easy to define homophily. Let

$$p = \frac{p_s + (m-1)p_d}{m} \quad (2)$$

be the average linking probability in the islands model (i.e., the probability that two agents drawn uniformly at random are linked).

One natural measure of homophily then compares the difference between same and different linking probabilities to the average linking probability, with a normalization of dividing by the number of islands, m :

$$h^{\text{islands}}(m, p_s, p_d) = \frac{p_s - p_d}{mp}. \quad (3)$$

Note that this is equivalent to Coleman’s (1958) homophily index specialized to the islands model:

$$\frac{\frac{p_s}{mp} - \frac{1}{m}}{1 - \frac{1}{m}}.$$

This is a measure of how much a group’s fraction of same-type links ($\frac{p_s}{mp}$) exceeds its population share ($1/m$), compared to how big this difference could be ($1 - 1/m$).¹²

The measure h^{islands} captures how much more probable a link to a node of one’s own type is than a link to a node of any other type, and varies between 0 and 1, presuming that $p_s \geq p_d$. If a node only links to same-type nodes (so that $p_d = 0$), then the average linking probability p becomes p_s/m and so $h^{\text{islands}} = 1$, while if nodes do not pay attention to type when linking, then $p_s = p_d$ and $h^{\text{islands}} = 0$. Indeed, the purpose of the m in the denominator is to restrict the variation of the measure exactly to the interval $[0, 1]$ (under the assumption that $p_d \leq p_s$).

¹²To see this, note that Coleman’s index can be rewritten as $\frac{p_s - p}{(m-1)p}$, and then from (2) it follows that $p_s - p = \frac{m-1}{m}(p_s - p_d)$; substituting verifies the equivalence.

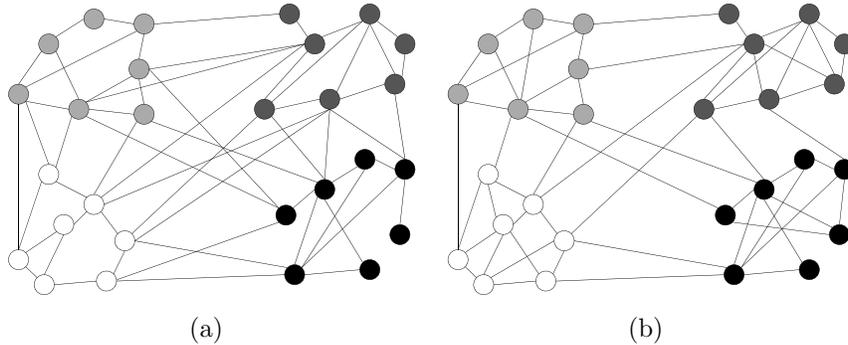


Figure 1: Islands networks with low and high homophily are shown in (a) and (b), respectively.

This simple measure of homophily introduced above is equal to the spectral homophily, as shown in Golub and Jackson (2012a).

PROPOSITION 1. If (\mathbf{P}, \mathbf{n}) is an islands network with parameters (m, p_s, p_d) , then

$$h^{\text{islands}}(m, p_s, p_d) = h^{\text{spec}}(\mathbf{P}, \mathbf{n}).$$

2.4 Average-Based Updating and its Speed

The processes that we focus on are ones where agents' behaviors or beliefs depend on some average of their neighbors' behaviors or beliefs. Applications include ones where agents dynamically and myopically best respond, trying to match the average behavior of their neighbors, as in common specifications of peer effects models¹³; the specification is also used in models of belief updating, as discussed in the introduction.

2.4.1 Definition

Average-based updating processes are described as follows. Given a network \mathbf{A} , let the matrix $\mathbf{T}(\mathbf{A})$ be defined by $T_{ij}(\mathbf{A}) = A_{ij}/d_i(\mathbf{A})$. Beginning with an initial vector of behaviors or beliefs $\mathbf{b}(0) \in [0, 1]^n$, agent i 's choice at date $t \geq 1$ is simply

$$b_i(t) = \sum_j T_{ij}(\mathbf{A})b_j(t-1).$$

¹³For recent examples in a network context, see Calvó-Armengol, Patacchini, and Zenou (2009) and Bramoullé, Djebbari, and Fortin (2009).

That is, the agent matches the average of his or her neighbors' last-period choices. In matrix form, this is written as:

$$\mathbf{b}(t) = \mathbf{T}(\mathbf{A})\mathbf{b}(t - 1)$$

for $t \geq 1$. It follows that

$$\mathbf{b}(t) = \mathbf{T}(\mathbf{A})^t \mathbf{b}(0).$$

2.4.2 Convergence

As long as the network is connected and satisfies a technical condition, the process will converge to a limit.¹⁴ In the results below about random networks, the assumptions ensure that the networks satisfy the conditions of the lemma with a probability tending to one as n grows.¹⁵

LEMMA 1. If \mathbf{A} is connected and has at least one cycle of odd length¹⁶, then $\mathbf{T}(\mathbf{A})^t$ converges to a limit $\mathbf{T}(\mathbf{A})^\infty$ such that $(\mathbf{T}(\mathbf{A})^\infty)_{ij} = \frac{d_j(\mathbf{A})}{D(\mathbf{A})}$.

2.4.3 Measuring Speed: Consensus Time

Consider some network \mathbf{A} and the linear updating process with the matrix $\mathbf{T}(\mathbf{A})$. To measure how fast average-based updating processes converge, we simply examine how many periods are needed for the vector describing all agents' behaviors or beliefs to get within some distance ε of its limit. The measure of deviation from consensus we use has a simple interpretation. At each period of updating, there are twice as many "messages" sent as there are links in the network – two messages across each link about each agent's behavior or belief. Let $\mathbf{m}(t)$ be this vector of messages for some ordering of the directed links. We define the distance from consensus at time t to be the root mean square distance of $\mathbf{m}(t)$ from its limit $\mathbf{m}(\infty)$. For the network \mathbf{A} and starting beliefs \mathbf{b} , we denote this distance (the *consensus distance*) at time t by $\text{CD}(t; \mathbf{A}, \mathbf{b})$.

¹⁴If the communication network is directed, then convergence requires some aperiodicity in the cycles of the network and works with a different segmentation into components, but still holds quite generally, as discussed in Golub and Jackson (2010).

¹⁵For example, Theorem 1 shows that under these assumptions, agents converge to within a specified distance of consensus beliefs in the worst case, which could not happen if the network were not connected.

¹⁶This is trivially satisfied if there is at least one i with $A_{ii} > 0$.

This measurement of distance between messages corresponds to a weighted version of the usual ℓ_2 norm. More specifically, given two vectors of beliefs \mathbf{v} and \mathbf{u} and a vector of weights \mathbf{w} , define $\|\mathbf{v} - \mathbf{u}\|_{\mathbf{w}} = [\sum_i w_i (v_i - u_i)^2]^{1/2}$. The distance of beliefs at time t from consensus is then $\text{CD}(t; \mathbf{A}, \mathbf{b}) = \|\mathbf{T}(\mathbf{A})^t \mathbf{b} - \mathbf{T}(\mathbf{A})^\infty \mathbf{b}\|_{\mathbf{s}(\mathbf{A})}$, where we use the weights $\mathbf{s}(\mathbf{A})$ defined by $\mathbf{s}(\mathbf{A}) = \left(\frac{d_1(\mathbf{A})}{D(\mathbf{A})}, \dots, \frac{d_n(\mathbf{A})}{D(\mathbf{A})}\right)$. In other words, $\|\mathbf{T}(\mathbf{A})^t \mathbf{b} - \mathbf{T}(\mathbf{A})^\infty \mathbf{b}\|_{\mathbf{s}(\mathbf{A})}^2$ is a weighted sum of differences between current beliefs and eventual beliefs, with agent i 's term weighted by his or her relative degree. This is equivalent to the ‘‘messages’’ interpretation because an agent with degree $d_i(\mathbf{A})$ sends a share of the messages in the network given exactly by $s_i(\mathbf{A}) = \frac{d_i(\mathbf{A})}{D(\mathbf{A})}$.

Then the consensus time is the time it takes for this distance to get below ε , for the worst-case initial beliefs¹⁷:

DEFINITION 2. The *consensus time* to $\varepsilon > 0$ of a connected network \mathbf{A} is

$$\text{CT}(\varepsilon; \mathbf{A}) = \sup_{\mathbf{b} \in [0,1]^n} \min\{t : \text{CD}(t; \mathbf{A}, \mathbf{b}) < \varepsilon\}.$$

3 The Theory Being Tested

A few definitions and notations are needed before stating the main theoretical result that we are testing. Throughout the section, we consider sequences of multi-type random networks, with all quantities (e.g., the matrix of intergroup linking probabilities \mathbf{P} and the vector of group sizes \mathbf{n}) indexed by the overall population size, n . We generally omit the explicit indexing by n to avoid clutter.

The next definition catalogs several regularity conditions on a sequence of multi-type random networks which will be assumed in the theorem.

DEFINITION 3.

1. A sequence of multi-type random networks is *sufficiently dense* if the ratio of the minimum expected degree to $\log^2 n$ tends to infinity. That is:

$$\frac{\min_k d_k(\mathbf{Q}(\mathbf{P}, \mathbf{n}))}{\log^2 n} \rightarrow \infty.$$

2. A sequence of multi-type random networks has *no vanishing groups* if

$$\liminf_n \frac{n_k}{n} > 0.$$

¹⁷Golub and Jackson (2012a) discusses the particulars of this definition in detail.

3. A sequence of multi-type random networks has *interior homophily* if

$$0 < \liminf_n h^{\text{spec}}(\mathbf{P}, \mathbf{n}) \leq \limsup_n h^{\text{spec}}(\mathbf{P}, \mathbf{n}) < 1.$$

4. Let \underline{P} denote the smallest nonzero entry of \mathbf{P} and \overline{P} denote the largest nonzero entry. A sequence of multi-type random networks has *comparable densities* if:

$$0 < \liminf_n \overline{P}/\underline{P} \leq \limsup_n \overline{P}/\underline{P} < \infty.$$

The sufficient-density condition ensures that with a probability tending to one, all nodes will be path-connected to each other.¹⁸ The no-vanishing-groups condition ensures that all groups have a positive share of the population. The interior-homophily condition requires that homophily not grow arbitrarily large or approach 0.¹⁹ Finally, the comparable-densities condition ensures that positive interaction probabilities do not diverge arbitrarily: they may be quite different, but their ratios must remain bounded.

The next definition is simply used to state the theorem's conclusion compactly:

DEFINITION 4. Given two sequences of random variables $x(n)$ and $y(n)$, we write $x(n) \approx y(n)$ to denote that for any $\varepsilon > 0$, if n is large enough, then

$$(2 + \varepsilon)^{-1}y(n) \leq x(n) \leq (2 + \varepsilon)y(n)$$

holds with probability at least $1 - \varepsilon$.

In other words, $x(n) \approx y(n)$ indicates that the two (random) expressions $x(n)$ and $y(n)$ are within a factor of 2 (with a vanishingly small amount of slack) for large enough n with a probability going to 1.²⁰

With these preliminaries, we can state a result from Golub and Jackson (2012a):

¹⁸The minimum needed for the network to be connected with high probability is for the degrees to grow faster than $\log n$. The condition here is a little stronger than this, and turns out to be what is needed to prove the tight asymptotic characterizations of convergence time we are about to present.

¹⁹The case of no homophily is dealt with in detail in Chung et al. (2004), and the case of homophily approaching 1 may lead the network to become disconnected, in which case there can be no convergence at all.

²⁰Note that the relation \approx is symmetric. In Golub and Jackson (2012a), we used a slightly different definition. None of the results of that paper depend on which definition is used.

THEOREM 1. Consider a sequence of multi-type random networks satisfying the conditions in Definition 3. Then, for any $\gamma > 0$:

$$\text{CT} \left(\frac{\gamma}{n}; \mathbf{A}(\mathbf{P}, \mathbf{n}) \right) \approx \frac{\log(n)}{\log(1/|h^{\text{spec}}(\mathbf{P}, \mathbf{n})|)}.$$

This result says that the speed of convergence of an average-based updating process is approximately proportional to $\log(1/|h^{\text{spec}}(\mathbf{P}, \mathbf{n})|)$. Moreover, the speed of the process essentially depends only on population size and homophily. The approximation for consensus time on the right-hand side is always within a factor of two of the true consensus time. Properties of the network other than spectral homophily can change the consensus time by at most a factor of two relative to the prediction made based on spectral homophily alone.

In the case of the islands model, the result has a simple corollary in view of Proposition 1 in Section 2.3:

COROLLARY 1. Consider a sequence of islands networks satisfying the conditions in Definition 3. Then, for any $\gamma > 0$:

$$\text{CT} \left(\frac{\gamma}{n}; \mathbf{A}(\mathbf{P}, \mathbf{n}) \right) \approx \frac{\log(n)}{\log(1/|h^{\text{islands}}(m, p_s, p_d)|)} = \frac{\log(n)}{\log \left| \frac{mp}{p_s - p_d} \right|}.$$

4 The Empirical Analysis

With theoretical predictions of consensus time based on the multi-type random networks in hand, we now compare those predictions to consensus time when the process is simulated on actual networks. In particular, we examine consensus times in 84 social networks from the Adolescent Health dataset and show how the patterns in those data illustrate our conclusions.

4.1 Data Description

The network data are from the in-school questionnaire from the 1994–95 Wave I of the Add Health study. Overall, 90,118 students took this survey. The particular data analyzed here come from a subsample of the 84 US high schools²¹ that had at least a fifty percent response rate. The overall sample of schools was chosen by systematic

²¹In some cases the high schools are paired with a “sister” middle school that was in the same community, and those students are also included in the networks.

sampling methods and implicit stratification, with the goal of making the sample representative of US schools with respect to region of country, urbanicity, school size, school type, and ethnicity.²²

The data include information regarding each student’s grade, gender and race. Grade is the year in school, ranging from 6 to 12. Race is self-reported as Asian, black, Hispanic, white, or other (with occasional missing data). Sex is self-reported as male or female. In addition, each student was asked to name his or her closest male and female friends.²³ Using the reported friendship networks (linking two individuals if either named the other as friend), yields the social networks. For each community, our analysis focuses on the largest connected component of the network.

We remove two networks due to convergence issues (as discussed in detail below). Ultimately, our analysis was performed on 82 connected networks, containing a total of 48,065 students (an average of 586 students per network, with a standard deviation of 426).

4.2 What Can the Data Tell Us?

Given the demographic data, we can compute homophilies for various definitions of types. We measure consensus times by simulating the DeGroot process and directly computing them in the sampled social networks.²⁴ In view of this, it is worth re-emphasizing the purpose of the empirical exercise. There are actually three aspects to the full spectrum of the results in Golub and Jackson (2012a) that could hypothetically be tested, depending on available data.

- First, there is a question of how well the multi-type random graph model fits

²²Much more information about the Add Health study design can be found in Harris et al. (2009).

²³The number of friends reported was capped at five of each type, or ten in total. Less than ten percent of the students hit the caps, but that still censors the data. This design feature makes homophilies computed based on sex somewhat less reliable than the others, since it would tend to equalize the numbers of reported male and female friends, even if there were strong homophily present.

²⁴For any given vector of initial beliefs, we can directly calculate how long it takes for beliefs to converge to within a given distance of consensus when the DeGroot process occurs on a given real network. In principle, then, to compute consensus time, one simply needs to search over starting beliefs. However, the space of potential starting beliefs is very large. Fortunately, it is possible to use certain eigenvalues and eigenvectors of the updating matrix $\mathbf{T}(\mathbf{A})$ to obtain theoretical bounds on the true consensus time and to make good guesses about which initial belief vectors will make the time to converge as large as possible (for the theoretical results along these lines, see the proof of Lemma 2 in Golub and Jackson, 2012a). With the guesses in hand as starting points, we then do Monte Carlo simulation to estimate the true consensus time.

the data when only a few simple observed types are used. The theoretical results are guaranteed to be verified, at least asymptotically, *if* the real high school networks were actually generated from the multi-type random graph model based on these simple types. However, the results might not provide good approximations otherwise. Thus, we can test how well implementations of the multi-type random graph model perform even when only using a few basic “types”.

- Second, the results quoted above are asymptotic and there is a question about whether our bounds on how consensus time relates to homophily will be useful in finite samples of medium size.
- Third, there is a question of whether or not people actually update their beliefs or behaviors in a way well-approximated by the average-based updating model.

Our empirical analysis answers the first two questions in the affirmative, and does not address the third question at all. In particular, the multi-type random network model is a good fit for these social networks when it comes to investigating consensus time, and gets a good deal of explanatory power from very basic definitions of types. Our claim that the study of the convergence of linear updating (or Markov) processes on large social networks can be reduced to simple computations about homophily is not merely an asymptotic, theoretical claim, but one that holds up well when applied to the data.

Whether or not these models of updating and communication shed light on actual social behavior – that is, on how people actually communicate and choose in networks – is obviously an important question, but one that requires additional (longitudinal) data and is left for future investigations.

4.3 Testing the Predictions

Our first step is to test the relationship in Theorem 1 between convergence and homophily. We begin by rearranging the expressions in the theorem into an (inverted) form that compensates for the extreme behavior of the convergence times at high homophilies and makes the quantities amenable to linear regressions.

In particular, let

$$\rho(X) = \exp\left(-\frac{\log n}{X}\right).$$

By Theorem 1, under the multi-type random network model, the quantity $\rho(\text{CT}(\gamma/n; \mathbf{A}))$ – which can be thought of loosely as an implied per-step rate of convergence – is approximately $h^{\text{spec}}(\mathbf{P}, \mathbf{n})$. To address the main empirical question of the paper, we run regressions of $\rho(\text{CT}(\gamma/n; \mathbf{A}))$ on the empirical analogue of $h^{\text{spec}}(\mathbf{P}, \mathbf{n})$ to see if the theoretical relationship holds up in the data. The empirical analogue of $h^{\text{spec}}(\mathbf{P}, \mathbf{n})$ is the second-largest eigenvalue of $\mathbf{F}(\mathbf{P}, \mathbf{n})$, with $Q_{k\ell}$ in the definition (equation 1) replaced by the actual total contribution to the sum of degrees of group k arising from links with group ℓ . (When k and ℓ are different, this is simply the number of links between groups k and ℓ .) In defining types for this exercise, we begin with the finest definition available in the data. Thus, we consider a “type” to be a specific combination of race, grade, and sex: so, for instance, a type would be all female Asians in grade 9. Thus, in a high school with two sexes, four races, and four grades, there are 32 types.

The regressions we present include an intercept term. This is not in Theorem 1. However, it turns out that if there is additional homophily within each type on dimensions not reported in the data, then there will be an intercept term in the model. Other types of mismeasurement can cause the slope on the homophily regressor to be different from 1. Some details regarding what biases we might expect from using imperfect type specifications are worked out for the case of the islands model in Section 4.4.

Table 1: Dependent variable = $\rho(\text{CT}(\gamma/n; \mathbf{A}))$

($N = 82$)	
Variable	Coefficient (t -statistic)
Intercept	0.560 (13.9)
h^{spec}	0.431 (9.41)
R^2	0.525

Lastly, we removed two data points whose consensus times exceeded our algorithms’ capacity. These networks (schools number 53 and 57) had consensus times on the order of several thousand, and computing them precisely was infeasible. So, from now on, we work with the 82 data points excluding those schools. Throughout the analysis, we set $\gamma = \sqrt{0.1} \approx 0.32$ in our computations of $\text{CT}(\gamma/n; \mathbf{A})$, which cor-

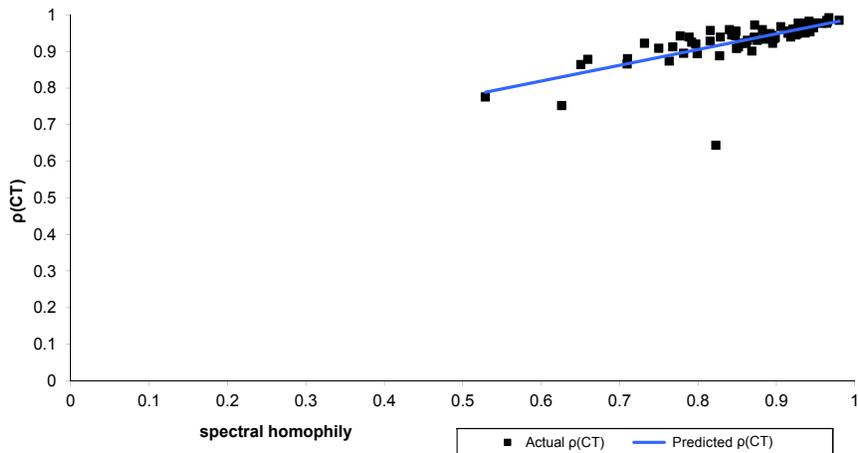


Figure 2: The relationship between $\rho(\text{CT}(\gamma/n; \mathbf{A}))$ and $h^{\text{spec}}(\mathbf{P}, \mathbf{n})$ computed based on the finest-grained type data available (i.e. a type is a grade-sex-race tuple).

responds to the mean squared deviation of messages from consensus being 0.1. The analysis does not appear to be sensitive to reasonable variations of this parameter.

The results of the regression are presented in Table 1 and Figure 2. The R^2 shows that the spectral homophily among these types accounts for roughly half of the variation in consensus times in the data. This is reasonably high in view of the fact that many type characteristics that potentially determine network formation – such as interests, extracurricular activities, socio-economic background, etc. – are not included in these data.

Spectral homophily based solely on grade-sex-race types has the advantage of explaining a large share of the variation in consensus times. However, since its definition involves an eigenvalue of a matrix that might be reasonably large (e.g., 32×32), it is worth asking whether we can predict consensus times using a simpler, more hands-on quantity.

Thus, from now on we work in the context of the islands model (recall Section 2.3). When all types have equal sizes, h^{spec} in Theorem 1 can be replaced by h^{islands} , as done in Corollary 1. The latter homophily is defined in terms of simple quantities – namely p_s , p , and the number of islands.²⁵ The probabilities have obvious empirical

²⁵As discussed in Section 2.3, we can write h^{islands} as $(p_s/p - 1)/(m - 1)$.

analogues: the observed density of links²⁶ within a type and overall.

Table 2: Dependent variable = $\rho(\text{CT}(\gamma/n; \mathbf{A}))$

($N = 82$)	
Variable	Coefficient (t -statistic)
Intercept	0.870 (61.2)
h^{islands} for full type	0.297 (4.91)
R^2	0.231

We note an important caveat. In the high school data, the types (whether defined by grade, sex, race, or combinations of these) will typically not have equal sizes. However, it turns out that by using the h^{islands} formula despite this imperfect correspondence of the data with the islands model, we can often obtain reasonably good predictions of consensus time. Given that h^{islands} has an advantage over h^{spec} in being simpler and more interpretable, we present results for a specification in which we regress $\rho(\text{CT}(\gamma/n; \mathbf{A}))$ on the empirical analogue of h^{islands} .

We present the results in Table 2 and Figure 3. In computing islands homophily in this regression, we use the finest definition of type available in the data in calculating the empirical p_s . That is, two nodes are considered to be the same type if they have the same grade, sex, and race. To set the total number of types, m , we assume that there are two equally-sized racial groups for each grade-sex combination. Thus, in a school that has both sexes and four grades, there would be 16 islands. While this is not typically exactly equal to the true number of distinct grade-sex-race tuples in the data, the largest two racial groups are usually substantially larger than the others, so it is a reasonable specification given that the islands model takes each type to be of equal size.

We also explore how much of the variation in consensus times can be explained by even simpler definitions of types. For example, out of the three demographic characteristics captured in the data, homophily based on grade has the greatest variation. The grades also have approximately equal sizes in most of the schools, so that it is particularly reasonable to use the formula from the equal-sized islands model. Thus,

²⁶That is, the number of actual links over the number of possible links.

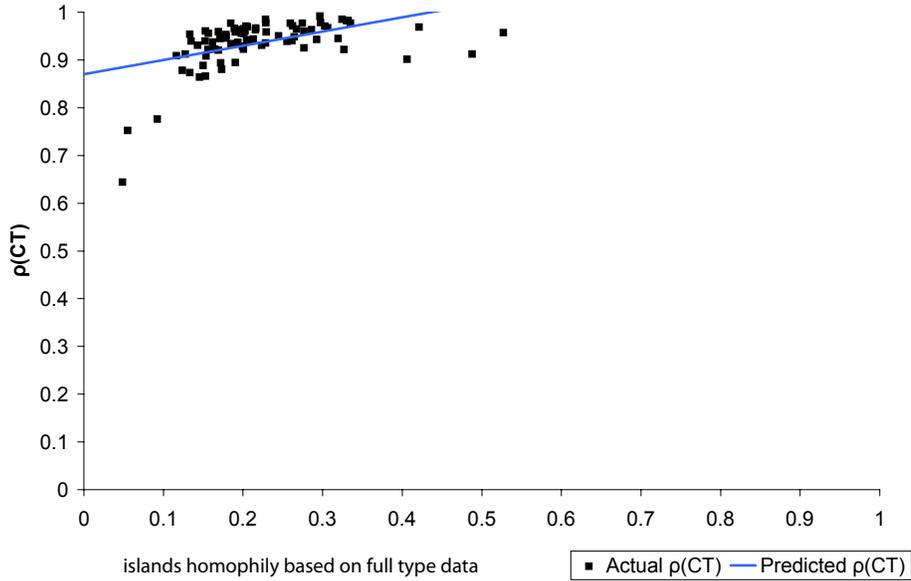


Figure 3: The relationship between $\rho(\text{CT}(\gamma/n; \mathbf{A}))$ and h^{islands} computed based on the finest-grained type data available (i.e. a type is a race-grade-sex tuple).

we regress $\rho(\text{CT}(\gamma/n; \mathbf{A}))$ on islands homophily when types are defined by grade in school alone. The results are reported in Table 3 and Figure 4.

Table 3: Dependent variable = $\rho(\text{CT}(\gamma/n; \mathbf{A}))$

($N = 82$)	
Variable	Coefficient (t -statistic)
Intercept	0.809 (32.6)
h^{islands} for grade	0.209 (5.21)
R^2	0.253

The fit is similar in quality to that obtained from the finest definitions of type under the islands model. Thus, under very coarse type specifications, consensus time estimations from the multi-type random network model capture significant amounts of variation in consensus time on actual networks.

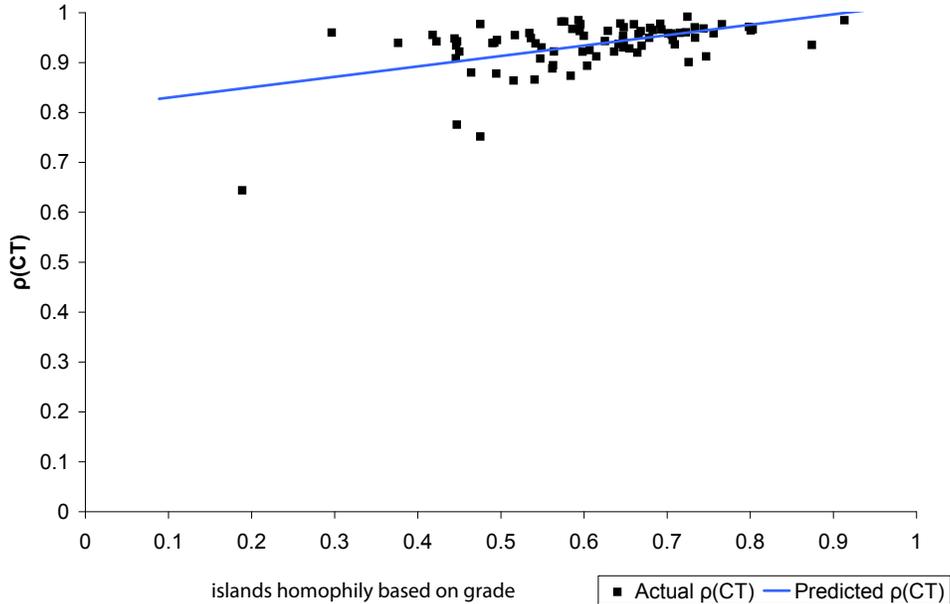


Figure 4: The relationship between $\rho(\text{CT}(\gamma/n; \mathbf{A}))$ and h^{islands} when types are defined based on grade in school only.

4.4 Imperfect Type Classifications

While the regressions show that spectral homophily plays a large role in predicting simulated consensus times on real networks, there are some ways in which the regressions depart from the predictions of the theory. According to Theorem 1, the regressions should have an intercept of zero, but a nonzero intercept term is quite clearly present. Moreover, the slope coefficient in each regression should be between $1/2$ and 2 ; our regressions show the actual slope coefficient being on the low side of that range.

We now show that such a discrepancy would be expected if the definition of types were leaving some attributes out. To this end, we examine an enriched model in which the econometrician imperfectly observes the types governing network formation.

Consider the following elaboration of the islands model. We have m equally sized islands N_1, \dots, N_m and each such island k is divided into r equally sized sub-islands N_{k1}, \dots, N_{kr} . If i and j are in different islands, then the linking probability between them is p_d . If i and j are in the same island, then the linking probability depends on whether they are in the same sub-island, or in different sub-islands. If they are in the

same sub-island, then they are linked with probability p_s . And if they are in the same island but different sub-islands, then they are linked with probability p_b . We assume $0 < p_d \leq p_b \leq p_s$, and that all link realizations are independent given the types of the nodes involved (i.e., this is a special case of the multi-type random network model).

The idea is that the researcher has data on the islands but not the sub-islands. For example, the special case where $p_b = p_d$ is one where the researcher has simply failed to disaggregate islands to the correct level.

We now show what happens if the homophily h is estimated as if the data were generated by the islands model with islands N_1, \dots, N_m and no sub-islands, despite the fact that there really are sub-islands.

If, without knowing about the sub-islands, we estimate the probability of same-type nodes being linked, we are actually estimating the quantity

$$\tilde{p}_s = \frac{p_s + (r - 1)p_b}{r},$$

and our estimate of h will be

$$\tilde{h} = \frac{\tilde{p}_s - p_d}{mp},$$

where p and p_d are presumed to be estimated correctly.

In this setting, it is not valid to apply Corollary 1 with the homophily computed based on the misidentified island structure. Instead, $h^{\text{islands}}(m, p_s, p_d)$ should be replaced by a consistent estimate of the spectral homophily $h^{\text{spec}}(\mathbf{P}, \mathbf{n})$. This spectral homophily will be an affine function of \tilde{h} , and this is the content of the following proposition.

PROPOSITION 2. In the modified islands setting just described, if \tilde{h} is the regressor given above (computed without information about the sub-island structure), then

$$h^{\text{spec}}(\mathbf{P}, \mathbf{n}) = \alpha + \tilde{h},$$

where α is independent of p_s and p_b .

The proof of this result appears in an appendix, Section 6. In running the regressions, we do not make use of the details of the formula for α derived in the proof. We merely note that there is an affine bias if there is some homophily inside the islands on dimensions falling outside the scope of our data. Thus, including an intercept in the regression of convergence rates on \tilde{h} is a reasonable first-order approximation to account for some of this affine bias.

Another realistic situation is one in which the econometrician’s notion of the island structure does not perfectly track the true islands N_1, N_2, \dots, N_m , but instead involves (i) grouping together some agents who are not actually in the same island for the purposes of network formation and (ii) separating others who should in fact be grouped together. In this case, we conjecture that, in regressions, one could obtain slope coefficients (on the mismeasured homophily \tilde{h}) not equal to one. Essentially, this is because in computing \tilde{h} , some truly intragroup density is classified as intergroup, and vice versa. The direction and magnitude of the distortion in the coefficient on \tilde{h} depends on the extent of each misclassification error. This might also explain why there appears to be some nonlinearity in the data in Figure 3 when we use our finest definition of types.

Of course, in more realistic settings, the various kinds of symmetry present in this simple illustration with the islands model will not exist. However, it appears that more general formulas or characterizations could be obtained describing how homophilies at various levels interact. This could be a useful direction to pursue in taking this model to other empirical settings, where there will almost always be some underlying homophily on dimensions not captured, or captured incorrectly, in the data.

5 Concluding Remarks

Various models of the structure of a network have been used as foundations to analyze how social network properties affect behavior. Here we have initiated an examination of whether a simulated process on actual networks matches predictions that are made assuming the network structure model is correct. This opens new questions for future research.

The first question is closely related to our technical results. An equivalent way of stating the result of this paper is that we can predict DeGroot process consensus times for actual networks by using the consensus time of an analogous process running on a highly simplified network which takes into account only the large-scale, group-level network properties.²⁷ To what extent does such a correspondence hold for properties *other* than consensus time, such as the expected disagreement between two particular groups in society after a certain number of periods of updating? In Section IV of Golub and Jackson (2012a), we show that, at least asymptotically as the network

²⁷Section IV of Golub and Jackson (2012a) formalizes this statement.

grows large, the correspondence should hold for many other properties. Testing this class of predictions using an approach similar to the one we have taken here presents an avenue for future work.

More broadly, one can ask how well other probabilistic models of network structure perform in similar exercises. How does the answer change as we keep the network model the same but vary the process occurring on the network? Which models of behavior mimic actual behaviors – at least when it comes to important aggregate statistics – and how does the answer to that question depend on the types of networks in which the agents interact?

6 Appendix: Proof of Proposition 2

Here we prove Proposition 2.

Proof of Proposition 2: Letting \mathbf{E}_k denote the matrix of all ones of size k and \mathbf{I}_k denote the identity matrix of size k , we find that with \mathbf{P} specified by the description above,

$$\mathbf{F}(\mathbf{P}, \mathbf{n}) = \frac{p_d \mathbf{E}_{mr} + (p_b - p_d) \mathbf{I}_m \otimes \mathbf{E}_r + (p_s - (p_b - p_d)) \mathbf{I}_{mr}}{p_s + (r - 1)p_b + (m - 1)rp_d}.$$

Now, the second eigenvalue of

$$p_d \mathbf{E}_{mr} + (p_b - p_d) \mathbf{I}_m \otimes \mathbf{E}_r$$

is $r(p_b - p_d)$. Thus,

$$\lambda_2(\mathbf{F}(\mathbf{P}, \mathbf{n})) = \frac{r(p_b - p_d) + p_s - (p_b - p_d)}{p_s + (r - 1)p_b + (m - 1)rp_d}.$$

This can be rewritten as

$$\lambda_2(\mathbf{F}(\mathbf{P}, \mathbf{n})) = \frac{r\tilde{p}_s - (r - 1)p_d}{mrp}$$

or

$$\lambda_2(\mathbf{F}(\mathbf{P}, \mathbf{n})) = \tilde{h} + \frac{p_d}{mrp}.$$

This completes the argument. ■

References

- BRAMOULLÉ, Y., H. DJEBBARI, AND B. FORTIN (2009): “Identification of Peer Effects through Social Networks,” *Journal of Econometrics*, 150, 41–55.
- CALVÓ-ARMENGOL, A., E. PATACCHINI, AND Y. ZENOU (2009): “Peer Effects and Social Networks in Education,” *Review of Economic Studies*, 76, 1239–1267.
- CHANDRASEKHAR, A., H. LARREGUY, AND J. P. XANDRI (2010): “Testing Models of Social Learning on Networks: Evidence from a Lab Experiment in the Field,” Mimeo., MIT.
- CHUNG, F. AND L. LU (2002): “The Average Distance in Random Graphs with Given Expected Degrees,” *Proceedings of National Academy of Science*, 99, 15879–15882.
- CHUNG, F., L. LU, AND V. VU (2004): “The Spectra of Random Graphs with Given Expected Degrees,” *Internet Mathematics*, 1, 257–275.
- COLEMAN, J. (1958): “Relational Analysis: The Study of Social Organizations with Survey Methods,” *Human Organization*, 17, 28–36.
- COPIC, J., M. O. JACKSON, AND A. KIRMAN (2009): “Identifying Community Structures from Network Data,” *B.E. Press Journal of Theoretical Economics (Contributions)*, 9:1, Article 30.
- CORAZZINI, L., F. PAVESI, B. PETROVICH, AND L. STANCA (2011): “Influential Listeners: An Experiment on Persuasion Bias In Social Networks,” CISEPS Research Paper No. 1/2011. Available at ssrn.com/abstract=1840565
- CURRARINI, S., M. O. JACKSON, AND P. PIN (2009): “An Economic Model of Friendship: Homophily, Minorities and Segregation,” *Econometrica*, 77, 1003–1045.
- DEGROOT, M. H. (1974): “Reaching a Consensus,” *Journal of the American Statistical Association*, 69, 118–121.
- DEMARZO, P., D. VAYANOS, AND J. ZWIEBEL (2003): “Persuasion Bias, Social Influence, and Uni-Dimensional Opinions,” *Quarterly Journal of Economics*, 118, 909–968.

- DIACONIS, P. AND D. FREEDMAN (1981): “On the Statistics of Vision: The Julesz Conjecture,” *Journal of Mathematical Psychology*, 24, 112–138.
- FIENBERG, S. E. AND S. WASSERMAN (1981): “An Exponential Family of Probability Distributions for Directed Graphs: Comment,” *Journal of the American Statistical Association*, 76, 54–57.
- FRENCH, J. (1956): “A Formal Theory of Social Power,” *Psychological Review*, 63, 181–194.
- GOLUB, B. AND M. O. JACKSON (2010): “Naïve Learning in Social Networks: Convergence, Influence, and the Wisdom of Crowds,” *American Economic Journal: Microeconomics*, 2, 112–149.
- (2012a): “How Homophily Affects the Speed of Learning and Best-Response Dynamics,” *Quarterly Journal of Economics*, 127, forthcoming.
- (2012b): “Network Structure and the Speed of Learning: Measuring Homophily Based on its Consequences,” *Annals of Economics and Statistics*, forthcoming.
- HARARY, F. (1959): “A Criterion for Unanimity in French’s Theory of Social Power,” in *Studies in Social Power*, ed. by D. Cartwright, Ann Arbor: University of Michigan Press.
- HARRIS, K., C. HALPERN, E. WHITSEL, J. HUSSEY, J. TABOR, P. ENTZEL, AND J. UDRY (2009): “The National Longitudinal Study of Adolescent Health: Research Design,” Available at www.cpc.unc.edu/projects/addhealth/design
- HOLLAND, P. W., K. B. LASKEY, AND S. LEINHARDT (1983): “Stochastic Block-models: First Steps,” *Social Networks*, 5, 109–137.
- JACKSON, M. O. (2008): *Social and Economic Networks*, Princeton, N.J.: Princeton University Press.
- JADBABAIE, A., P. MOLAVI, A. SANDRONI, AND A. TAHBAZ-SALEHI (2012): “Non-Bayesian Social Learning,” *Games and Economic Behavior*, forthcoming, available at ssrn.com/abstract=1916109
- JADBABAIE, A. AND A. TAHBAZ-SALEHI (2010): “Consensus Over Ergodic Stationary Graph Processes,” *IEEE Transactions on Automatic Control*, 55.

- MARSDEN, P. V. (1987): “Core Discussion Networks of Americans,” *American Sociological Review*, 52, 122–131.
- MCPHERSON, M., L. SMITH-LOVIN, AND J. COOK (2001): “Birds of a Feather: Homophily in Social Networks,” *Annual Review of Sociology*, 27, 415–444.
- MCSHERRY, F. (2001): “Spectral Partitioning of Random Graphs,” in *Proceedings of the 33rd Annual ACM Symposium on Theory of Computing (STOC 2001)*.
- MUELLER-FRANK, M. (2012): “The Efficiency of Social Learning in Networks with Bayesian and Non-Bayesian Agents,” Mimeo., available at ssrn.com/abstract=2040352