A STRUCTURAL MODEL OF HOMOPHILY AND CLUSTERING IN SOCIAL NETWORKS

ANGELO MELE
JOHNS HOPKINS UNIVERSITY

ABSTRACT. Social networks display homophily and clustering, and are usually sparse. I develop and estimate a structural model of strategic network formation with heterogeneous players and latent community structure, whose equilibrium networks are sparse and exhibit homophily and clustering. Each player belongs to a community unobserved by the econometrician. Players' payoffs vary by community and depend on the composition of direct links and common neighbors, allowing preferences to have a bias for similar people. Players meet sequentially and decide whether to form bilateral links, after receiving a random matching shock. The probability of meeting people in different communities is smaller than the probability of meeting people in the same community, and it decreases with the size of the network. The model converges to a hierarchical exponential family random graph, with weak dependence among links. As a consequence the equilibrium networks are sparse and the sufficient statistics of the network concentrate around their mean. The posterior distribution of structural parameters and unobserved heterogeneity is estimated with school friendship network data from Add Health, using a Bayesian exchange algorithm. The estimates detect high levels of racial homophily, and heterogeneity in both costs of links and payoffs from common friends. The posterior predictions show that the model is able to replicate the homophily levels and the aggregate clustering of the observed network, in contrast with standard exponential family network models.

JEL CODES: C13, C31, C57, D85
Keywords: Social networks, exponential random graphs, weak dependence, homophily, clustering, sparse networks

1. INTRODUCTION

Social networks are important determinants of economic success. The number and composition of social relationships affect individual behavior and choices in different contexts: health, education, crime, investment, politics, employment, new product adoption and many others.[1] Therefore it is crucial to understand how these relationships are formed, what network architectures are optimal, and what policies can affect their shape.

This version: July 3, 2018. First version: August 20, 2017. Contact: angelo.mele@jhu.edu. I am grateful to Stephane Bonhomme, Vincent Boucher, Aureo dePaula, Michael Leung, Bryan Graham, Demian Pouzo, Lingjiong Zhu, Shuyang Sheng, Roger Moon, Zhongjun Qu and Hiro Kaido for comments and suggestions. Special thanks to Michael Schweinberger for helping with the use of his package `hergm`.

[1]See Jackson (2008), DePaula (forthcoming), Chandrasekhar (2016), Acemoglu et al. (2011), Golub and Jackson (2011), Fafchamps and Gubert (2007), Laschever (2009), Topa (2001), Conley and Udry (2010), Echenique and Fryer (2007), Nakajima (2007), De Giorgi et al. (2010), Goldsmith-Pinkham and Imbens (2013), Calvo-Armengol et al. (2009) for examples.

The analysis developed in this paper builds on three empirical observations about social networks. First, most observed networks are *sparse*; that is the number of links is proportional to the number of nodes.[2] This means that most individuals will not form too many links, even if the size of the network is very large. Second, social networks display *homophily*: people tend to have most interactions with similar individuals. This tendence has been shown for observable characteristics like race, gender, age, income, education and other demographics.[3] In addition, there may be homophily based on unobservables [4]. Third, observed social networks exhibits *clustering*: if two people have a link to a common neighbor, it is very likely that they are also connected to each other.[5] A useful structural model of network formation should be able to replicate these properties observed in the data.

I propose a structural model of network formation with heterogeneous players and latent community structure, that is able to match homophily, clustering and sparsity observed in social networks, as an equilibrium outcome. The model's equilibrium belongs to the class of discrete exponential family random graphs:[6] the main innovation is the introduction of a latent community structure to model unobserved heterogeneity in preferences and opportunities to form links. As a consequence, the model converges to a hierarchical exponential random graph model (Schweinberger and Handcock (2015)), where links are weakly dependent in equilibrium.

This approach has several advantages. First, the network formation model presented here generates sparse graphs. This is in contrast with the standard exponential family random graphs, that may generate dense networks.[7] Second, the model imposes weak dependence among links in equilibrium, avoiding the issue of degenerate probability distributions over networks, common in the ERGM literature[8] and in evolutionary game theory.[9] Third, it incorporates unobserved heterogeneity in the form of latent community structure, allowing me to separately identify unobserved heterogeneity, homophily and clustering in a tractable way.[10] Finally, for suitable parameterizations, it is possible to identify and estimate the structural parameters using only one network observation. This is in contrast with standard

[2]See Lovasz (2012), Chatterjee and Varadhan (2011), Jackson (2008), Chandrasekhar and Jackson (2014), Chandrasekhar (2016), DePaula (forthcoming) for formal definitions of sparse and dense graphs.

[3]See Jackson (2008), Currarini et al. (2009), Moody (2001), Mele (2017), DePaula et al. (forthcoming) and Boucher (2015) for examples.

[4]See Graham (2017), Dzemski (2017) , Boucher and Mourifie (forthcoming), Leung (2014).

[5]See Jackson (2008) and Jackson and Rogers (2007).

[6]Moody (2001), Snijders (2002), Caimo and Friel (2011), Boucher and Mourifie (forthcoming), Mele (2017)

[7]See for example the analyses in Diaconis and Chatterjee (2013), Mele (2017), Chatterjee and Varadhan (2011), Aristoff and Zhu (2014), Chandrasekhar and Jackson (2016).

[8]See Snijders (2002), Diaconis and Chatterjee (2013), Mele (2017), Chandrasekhar and Jackson (2014), Schweinberger and Handcock (2015) for examples.

[9]See Jackson and Watts (2001), Blume (1993) for examples.

[10]Previous work has introduced unobserved heterogeneity in network formation models. For example Graham (2017) and Dzemski (2017) include unobserved heterogeneity as additive in the payoffs, but do not include clustering in their specification. Leung (2014) and Boucher and Mourifie (forthcoming) include homophily based on unobserved heterogeneity. Breza et al. (2017) also include a community structure in a latent position model, but their model does not rely on a microfoundation of link formation, as in Mele (2017) or Badev (2013).

exponential family network models, which have been shown to have identification problems in the large network asymptotic framework.[11]

The network is formed sequentially: in each period two players are randomly selected from the population and meet. Upon meeting, players decide whether to update their link, by myopically maximizing the sum of their current utility. In the absence of any shock to the preferences, this process of network formation is consistent with pairwise stability with transfers, a common equilibrium notion used in the network formation literature.[12] I show that, conditional on the community structure, the network formation process can be characterized as a potential game and in the long-run the sequence of link updates converges to a stationary distribution over networks, which is a discrete exponential family with intractable normalizing constant. This implies that in the long run, the observed networks are pairwise stable (with transfers) with very high probability.[13] This result leverages the microeconomic foundations developed in Mele (2017).

The unobserved heterogeneity affects agents' preferences and how they meet to form links. The population of players is partitioned into non-overlapping communities. Upon birth, each player is randomly assigned to a community, that affects the probability of meeting other people in two ways: first, agents meet members of the same community more often than members of another community; second, the probability of meeting a member of another community decreases with the size of the network.

Preferences are defined over networks, covariates and community structure. The players' payoffs depend on the composition of direct connections, but also on the number of common friends. Preferences also depend on the unobserved heterogeneity: members of different communities are allowed to have different costs of forming links and different payoffs from common friends. I assume that players only care about common friends in the same community.[14] I prove that the latter assumption, and the decreasing probability of meeting members of other communities, generate a sparse network in equilibrium, where links are weakly dependent. As a consequence the sufficient statistics of the network are approximately normal for a large number of communities,[15] including the number of links, the aggregate homophily levels and the aggregate clustering. One important implication of this result in estimation is that the sufficient statistics of the model tend to concentrate around

[11]See Diaconis and Chatterjee (2013), Mele (2017), Aristoff and Zhu (2014) for discussions about the identification problems in exponential random graph models.

[12]See Jackson (2008) for a review of the equilibrium concepts in the theoretical literature on network formation in economics. Christakis et al. (2010)'s model is similar in spirit, but focuses on estimation and does not provide a characterization of the equilibria.

[13]See Monderer and Shapley (1996) for a general definition and discussion about potential games. Similar characterizations of equilibria can be found in Butts (2009), Jackson and Watts (2001), Badev (2013), Hsieh and Lee (2012).

[14]Similar truncated preferences are used in Jackson (2008) and DePaula et al. (forthcoming), among others.

[15]See Schweinberger and Handcock (2015) for a discussion. In principle, one could apply standard regularity conditions to show that the maximum likelihood estimator is asymptotically normal for suitable parameterizations. I rely on Bayesian inference instead, following the approach of Schweinberger and Handcock (2015).

their expected value, while this is not necessarily the case with standard exponential family random graphs (Diaconis and Chatterjee (2013), Mele (2017), Aristoff and Zhu (2014)).

A major challenge for estimation is that the likelihood of the model is invariant to the labeling of the communities. That is, if we consider a permutation of the community labels, we will have the same likelihood. This problem is often encountered in the finite mixture models literature (McLachlan and Peel (2000), Stephens (2000)), and it causes an identification problem. The Bayesian approach developed in Schweinberger and Handcock (2015) is able to tackle this particular challenge. Since the community structure is unobserved by the econometrician, the empirical strategy is to impose a prior distribution over communities and parameters, and perform hierarchical Bayesian inference to recover the structural preference parameters. The community structure is assumed to follow an i.i.d. multinomial distribution, as it is standard in the literature on stochastic blockmodels (Airoldi et al. (2008)). A standard Bayesian approach would impose a Dirichelet prior on the multinomial parameters; however, such choice of the prior will make the posterior invariant to permutations of the community labels. Therefore, a nonparametric prior is preferred, along the lines of Ishwaran and James (2001), that is not invariant to permutations of the labels. Therefore the posterior is also not invariant with respect to labels.

To further reduce this problem, the output of the posterior simulation is relabeled, using an algorithm developed in Stephens (2000) for finite mixture models. Because of the labeling issue, during the posterior simulations the labels may switch several times across communities, producing a MCMC output that is unreliable for inference. The most reliable approach in the Bayesian literature on finite mixtures is to use an algorithm that relabels the output of the MCMC, picking a particular labeling order of the communities. These algorithms choose a particular labeling by minimizing a loss function (Stephens (2000), McLachlan and Peel (2000), Gelman et al. (2003)) and are therefore rooted in a Bayesian framework.[16] This allows me to interpret the community-specific parameters and to perform inference.

The data used in estimation contains the network of friendships in a US high school, extracted from *The National Longitudinal Study of Adolescent Health* (Add Health).[17] The data also contains race, gender, grade and parental income of each student. The empirical estimates show high levels of homophily: preferences are biased towards links with students of the same racial group, grade and (parental) income levels. Furthermore, different communities have different costs of linking and different payoffs from common friends. The latter result proves that it is important to include unobserved heterogeneity in empirical network formation models, as there could be unobserved characteristics that make agents more or less social on average.

[16] I use the algorithm of Schweinberger and Handcock (2015), that uses Simulated Annealing to minimize the loss function at each iteration.

[17] 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 Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 17 other agencies. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Persons interested in obtaining Data Files from Add Health should contact Add Health, The University of North Carolina at Chapel Hill, Carolina Population Center, 123 W. Franklin Street, Chapel Hill, NC 27516-2524 (addhealth@unc.edu). No direct support was received from grant P01-HD31921 for this analysis.

The model fits the data quite well. My simulations show that the posterior predictions of the estimated model are able to replicate the homophily levels and aggregate clustering of the network. As a comparison, standard exponential-family random graph models may not be able to match these network statistics, as they are often degenerate: that is, they put a very high probability mass on a few graphs, that are often close to the complete graph or the empty network.[18] The structural model developed here avoids degeneracy because of the weak dependence among links.[19]

This paper contributes to the literature on empirical network formation models by developing a structural model that is able to generate homophily and clustering as an equilibrium outcome in sparse networks. The models studied in Mele (2017) and Mele and Zhu (2017) are special cases of the model presented here, where there is only one community.[20] My model generates a hierarchical exponential random graph in equilibrium (Schweinberger and Handcock (2015)): this family of models imposes a community structure on the standard discrete exponential family random graphs (ERGMs), generating a network with weak dependence among links, whose likelihood can be factorized in between- and within-community components. I provide a game-theoretical counterpart of Schweinberger and Handcock (2015)'s statistical work, by leveraging the strategic approach developed in the economics literature on networks.[21] In fact, my structural model mantains the flexible specification of the exponential family random graphs, while incorporating the strategic and equilibrium microfoundations introduced in Mele (2017), and adding conditions that guarantee sparsity and weak dependence among links. I show that the equilibrium characterization using the theory of potential games and the long-run stationary behavior of the model shown in Mele (2017) survive the additional assumptions required for sparsity and limited dependence.

Unobserved heterogeneity has been modeled in different ways in the empirical network literature. Graham (2017) and Dzemski (2017) model unobserved heterogeneity as additive in the preference, for models of dyadic link formation, but they exclude payoffs from common neighbors in the utility function.[22] Boucher and Mourifie (forthcoming) and Leung (2014) show that limited dependence and sparsity are crucial to obtain consistency and asymptotic normality of estimators. While in their model the limited dependence is achieved by assuming that homophily must hold asymptotically, my model weakens the dependence among links through the community structure and the assumptions on the meeting process. Breza et al. (2017) also assume a latent community structure, but their model does not leverage the microfoundations in link formation provided here (conditional on the latent variables).

[18]See Snijders (2002), Chandrasekhar and Jackson (2014), Diaconis and Chatterjee (2013), Mele (2017) for a discussion.

[19]See the discussion in Schweinberger and Handcock (2015) and the Appendix B.

[20]Analogously, with the inclusion of additional payoffs for link externalities, this model includes the family of exponential random graphs as a special case with one community.

[21]See Jackson (2008), Jackson and Wolinsky (1996), Jackson and Watts (2001), Bala and Goyal (2000), Galeotti (2006).

[22]Similar approaches are in Charbonneau (2017), Jochmans (2017) and the literature on the β-model in statistics.

Structural models of network formation usually include biases in preferences or meetings to generate homophily in equilibrium (Currarini et al. (2009), Boucher (2015), Mayer and Puller (2008)); some also include payoffs from common neighbors (DePaula et al. (forthcoming), Menzel (2015), Sheng (2012), Ridder and Sheng (2015)). However, disentangling homophily, clustering and unobserved heterogeneity using only one network observation is extremely challenging (Graham (2017), Chandrasekhar and Jackson (2016)). My model contributes to the literature by separately identifying the payoffs for homophily and clustering, while modeling unobserved heterogeneity in a tractable way.

Previous work has noted the importance of sparsity for identification and the statistical properties of the estimators. Chandrasekhar and Jackson (2014) show that sparsity is one of the crucial ingredients to prove consistency of the estimates; DePaula et al. (forthcoming) show that sparsity and limited dependence lead to identification in a model of network formation with homophily.[23] The model in this paper has equilibrium networks that are sparse; the main assumption that generates this outcome is the asymptotically vanishing probability of meeting people in different communities. As in previous work sparsity is crucial to prove asymptotic normality of the sufficient statistics, because it reduces the dependence among links and allows me to focus the analysis on weakly dependent subnetworks.

The estimation of the model is computationally intensive, because the community structure is unobserved. However, one could pre-process the data and use an algorithm to identify and estimate community memberships before the estimation of the network model, like in Bonhomme et al. (2017). This would speed up computations and estimation, because the simulation of the community structure is the main computational bottleneck.[24] Some authors have partitioned the network into subgraphs to compute bounds for the identified sets (Sheng (2012)), while others have considered random graph models where subgraphs (rather than links) are formed (Chandrasekhar and Jackson (2016)). In this paper, I impose conditions that allow me to factorize the likelihood into within- and between-communities subgraphs components, and I implicitly focus on dependence within subnetworks, rather than across subnetworks, thus reducing the computational burden.

Finally, this paper is related to the study of structural models of complementary choices (Berry et al. (2014)) and models of demand for bundles (Gentzkow (2007)). There is a relationship between potential games and discrete choice models of bundles that was noted in Fox and Lazzati (forthcoming). With minimal modifications of the model presented here, one can interpret the community structure as limiting the dependence among different goods in a bundle, and excluding specific interactions among goods in the utility of the consumers to increase the speed of simulations. Similar models have recently been proposed in the marketing literature (Kosyakova et al. (2017)).

The rest of the paper is organized as follows. Section 2 presents the structural model and the equilibrium characterization, providing the main result on sparsity. Section 3 develops the estimation strategy and the asymptotic behavior of the sufficient statistics. Section 4

[23] Menzel (2015) generates sparsity assuming that the players' marginal cost of linking is the max of samples drawn from the Gumbel distribution, with the number of samples growing at a certain rate as the network grows.

[24] An alternative is the moderate use of parallelization for estimation of each community normalizing constants.

focuses on the empirical results, showing that friendship school networks exhibit high level of homophily and clustering, with a moderate level of unobserved heterogeneity in preferences. Section 5 concludes. The appendices contain additional theoretical results, proofs and details about the estimation.

2. A Structural Model of Network Formation

The economy consists of n players, each characterized by an M-dimensional vector $x_i = \{x_{i,1}, x_{i,2}, ..., x_{i,M}\}$ of observable characteristics. For example, x_i may include the gender, racial group, income, education levels and age of each individual in the economy.

Each player belongs to a *community*: this membership is observed by the players, but not by the econometrician. The community models unobserved heterogeneity in a very specific form: there exists some unobserved characteristics that separates individuals into different types. For example, some individuals are more social than others, a personality trait that is difficult to observe; some people care a lot about having a tightly-knit group of friends, others do not care as much. A player's type affects her preferences as well as the probability of meeting other people, as explained in detail below. Formally, the community structure is a partition of the n players in K subsets $\{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_K\}$. The K-dimensional vector of binary indicators $z_i = \{z_{i,1}, z_{i,2}, ..., z_{i,K}\}$ contains the membership information for player i. That is, player i belongs to community k if $z_{i,k} = 1$ and $z_{i,l} = 0$ for all the $l \neq k$. I will consider a model in which individuals can be member of one community only.[25] The researcher cannot observe the memberships vectors z_i, nor the number of communities K.

Before the network formation, nature chooses who belongs to each community (the community structure), according to a multinomial distribution. This is a relatively standard assumption in the stochastic blockmodels literature (Airoldi et al. (2008)).[26] I assume that the community assignment to each player is i.i.d.

$$Z_i|\eta_1, ..., \eta_K \overset{iid}{\sim} Multinomial\,(1; \eta_1, ...\eta_K) \text{ for } i = 1, ..., n \tag{1}$$

The $n \times M$ matrix x contains all the vectors of observable characteristics, and the $n \times K$ matrix z contains all the vectors of membership indicators.

The network of interactions is represented by a $n \times n$ matrix g, the adjacency matrix of the network, whose generic element $g_{ij} = 1$ if there is a link between i and j, and $g_{ij} = 0$ otherwise. I will consider an *undirected* network, with a symmetric matrix g. Some of the results below can be easily extended to directed networks.[27]

The network formation process works as follows. In period 0, nature randomly chooses the communities for each player i. Conditional, on the community structure $Z = z$, the network is formed sequentially as in Mele (2017) and Mele and Zhu (2017): in each period, two random players, i and j meet with probability $\rho(g, z_i, z_j, x_i, x_j, n)$. This probability can

[25] In principle, this community structure allows for multiple memberships. This is consistent with some statistical models in the literature, like the mixed membership stochastic blockmodel (MMSB) of Airoldi et al. (2008). However, it is not clear that one could identify preference for clustering in such a model.

[26] Schweinberger and Handcock (2015) assume an equivalent specification. The main advantage of the multinomial specification is that priors can be specified in a nonparametric way to speed up the computations of the posterior.

[27] See Mele (2017), Badev (2013), Schweinberger and Handcock (2015), for examples.

depend on the existing network g: for example, two people may meet more often if they have some friends in common. Furthermore, it is a function of the size of the network n: it could be easier to meet people in small networks than in large networks, for example. The function ρ also depends on the unobservable communities indicators z_i and z_j: people belonging to the same community may have more opportunities to meet than people in different communities, for example. Finally, it can also depend on the observable characteristics x_i and x_j: for example, people with similar observable characteristics may meet more often.[28]

Upon meeting i and j decide whether to cooperatively update their link g_{ij}: if the link does not exists, they decide whether to form the link; if the link already exists, they choose whether to delete the link. When choosing whether to update the link, players behave myopically, and do not consider how their decision affects the future evolution of the network.[29] I assume that players i and j maximize their joint payoff, when updating a link; this decision rule is compatible with pairwise stability with transfers, one of the most common equilibrium notions used in the network literature.[30]

2.1. **Meeting technology.** The probability $\rho(g, z_i, z_j, x_i, x_j, n)$ governs the opportunity to create and delete links. Observed networks are usually sparse, but contain some clusters of relatively more dense subnetworks(Jackson and Rogers (2007)). To replicate these empirical features, I impose the following assumptions on the function ρ. Let g_{-ij} denote the network g with the exclusion of link g_{ij}.

ASSUMPTION 1. *Conditional on the unobserved community structure z, the meeting are i.i.d. over time and the probability that i and j meet is*

$$\rho(g, z_i, z_j, x_i, x_j, n) = \begin{cases} \rho_w(g_{-ij}, x_i, x_j, n) & \text{if } z_i = z_j, \\ \frac{\rho_b(g_{-ij}, x_i, x_j)}{n^\delta} & \text{otherwise} \end{cases} \tag{2}$$

where $\rho_w(g_{-ij}, x_i, x_j, n)$ increases with n. For any n, the probabilities $0 \leq \rho_b(g_{ij}, x_i, x_j) \leq \rho_w(g_{ij}, x_i, x_j, n) \leq 1$ for any pair (i, j) and $\delta > 0$ is a scalar. I also assume that the sum of these probabilities over all possible pairs of players is one.

Assumption 1 states that players that belong to the same community can meet and form links at a rate $\rho_w(g_{-ij}, x_i, x_j, n) > 0$, which increases with the size of the network n; players in different communities meet at a rate that is decreasing with the size of the network $\rho_b(g_{-ij}, x_i, x_j)n^{-\delta}$. Both meeting probabilities may depend on the structure of the network g_{-ij}, and observable characteristics x_i and x_j. Notice that as n grows large, the probability

[28] See Currarini et al. (2009) and Currarini et al. (2010) for a model where meetings are biased in favor of people of the same group. Mele (2017), Badev (2013) and Chandrasekhar and Jackson (2014) also consider variants of this meeting technology.

[29] This modeling approach has been used in previous work by Nakajima (2007), Mele (2017), Mele and Zhu (2017), Badev (2013), Bala and Goyal (2000), Jackson and Watts (2001) among others.

[30] See Jackson (2008), Mele and Zhu (2017), Chandrasekhar and Jackson (2014) for examples.

that i meets someone in another community goes to zero at a rate that depends on the positive scalar δ. This is one of the ingredients that generates sparsity in the network.[31]

Assumption 1 models social interactions in a local way: as the network grows we may get opportunities to meet people outside our usual social circle, but we tend to keep most of the daily interactions local.

2.2. **Preferences.** Players' preferences are defined over networks, observables and community structures. Players receive payoffs from their direct connections, but also externalities from common friends.

Let $U_i(g, x, z; \theta)$ denote the utility of player i from network g, observable characteristics x, community structure z and parameters $\theta = \{\alpha, \beta, \gamma\}$. Preferences are described by

$$U_i(g, x, z; \theta) = \sum_{j=1}^{n} g_{ij} \left[u(x_i, x_j, z_i, z_j; \alpha, \beta) + \sum_{r \neq i,j}^{n} g_{jr} g_{ri} v(z_i, z_j, z_r; \gamma) \right] \tag{3}$$

where $u(x_i, x_j, z_i, z_j; \alpha, \beta)$ and $v(z_i, z_j, z_r; \gamma)$ are the payoffs of direct connections and common friends respectively.

Player i receives a direct payoff $u(x_i, x_j, z_i, z_j; \alpha, \beta)$ for each link she creates, that is when $g_{ij} = 1$. This payoff may depend on the unobservable communities z_i and z_j: for example, a person may have a bias for people in the same community; it may also depend on the observable characteristics x_i and x_j: for example, preferences may be a biased towards links with people of the same race, gender, income level, etc.

The payoff $u(x_i, x_j, z_i, z_j; \alpha, \beta)$ includes both costs and benefits of direct connections, so it should be interpreted as net benefit of forming a link. We assume as in Jackson and Wolinsky (1996) that players pay a cost for direct links, but not indirect connections. In this setting, α is the parameter that governs the cost of forming links, which is allowed to vary across communities; β is the parameter related to the benefits of forming links.

Player i receives an additional payoff $v(z_i, z_j, z_r; \gamma)$ for each friend in common with j. I allow the payoff to vary across communities. This captures the possibility that different communities value indirect connections in different ways.

Throughout the paper, the payoffs functional forms are restricted for tractability and identification purposes according to the following assumption.

ASSUMPTION 2. *The payoff from direct links $u(x_i, x_j, z_i, z_j; \alpha, \beta)$ is symmetric in observables x_i, x_j and latent community indicators z_i, z_j,*

$$u(x_i, x_j, z_i, z_j; \alpha, \beta) = u(x_j, x_i, z_j, z_i; \alpha, \beta) \tag{4}$$

[31] I show in appendix that the value of parameter δ is important for identification, estimation and the asymptotic properties of the estimators. See also Schweinberger and Handcock (2015) for an analogous definition of sparsity. They define networks as sparse when there exists constant $A > 0$ and $\delta > 0$ such that $\mathbb{E}(g_{ij}) \leq A\delta^{-n}$ for i and j that belong to different communities. I can prove that our restriction on the meeting process, together with the next few assumptions, implies their condition on sparsity. See appendix for details.

and the payoff from common neighbors $v(z_i, z_j, z_r; \gamma)$ is zero if players i, j and r belong to different communities:

$$v(z_i, z_j, z_r; \gamma) = \begin{cases} \gamma_k & \text{if } z_{ik} = z_{jk} = z_{rk} = 1 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

The simmetry in $u(x_i, x_j, z_i, z_j; \alpha, \beta)$ is required for identification in an undirected network. The indirect links payoff $v(z_i, z_j, z_r; \gamma)$ is nonzero only if all the individuals i, j, r are in the same community. This part of the assumption is crucial to obtain enough sparsity in the model, so that we are able to match the aggregate clustering of the observed network. If we allow the payoff of common friends to be nonzero across communities, we are implicitly assuming that the link between i and j depends on the existence and configuration of links among all the other players. This would impose a strong dependence in the network that would generate a dense graph, like in Mele (2017) or Mele and Zhu (2017). Assumption 2 generates weak dependence; adding Assumption 1, I can show that the dependence is asymptotically vanishing. Analogous restrictions have appeared in DePaula et al. (forthcoming), Jackson and Wolinsky (1996) and Menzel (2015), Leung (2014), Boucher and Mourifie (forthcoming).

Finally, I assume that players receive a joint matching shock to the preferences before choosing whether to update a link. The random shock models idiosynchratic reasons that could affect the decision to link: for example, in a particular period, a player i could be in a bad mood and reject a link to another player j that would have generated positive surplus.

ASSUMPTION 3. *Players receive a logistic matching shock before updating their links, which is i.i.d. over time and across pairs.*

Assumption 3 is standard in discrete choice models, and it is crucial to obtain the likelihood in closed-form:[32] I can therefore perform maximum likelihood or Bayesian estimation.[33]

2.3. **Equilibrium.** The structure imposed on this model by Assumptions 1, 2 and 3 generates sparse graphs, allowing for homophily in observables and unobservables, and clustering among players. The network formation model is a potential game (Monderer and Shapley (1996)): I can prove that there exists a potential function that summarizes the deterministic incentives of all the players.[34]

PROPOSITION 1. *Conditional on the community structure z, the network formation game is a potential game and there exists an aggregate potential function Q that summarizes*

[32] See also Mele (2017), Mele and Zhu (2017), Chandrasekhar and Jackson (2014), Heckman (1978).

[33] An alternative to assumptions 1-3 is to use the spatial GMM (Conley (1999), Conley and Topa (2007)) or the Approximate Bayesian Computation (ABC) that do not require to know the likelihood in closed-form (Marjoram et al. (2003), König (2016)).

[34] See also Mele and Zhu (2017), Butts (2009), Badev (2013), Hsieh and Lee (2012) and Chandrasekhar and Jackson (2014).

the incentives of the players to form links upon meeting

$$Q(g,x,z;\theta)=\sum_{i=1}^{n}\sum_{j=1}^{n}g_{ij}u(x_i,x_j,z_i,z_j;\alpha,\beta)+\frac{1}{6}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{r\neq i,j}^{n}g_{ij}g_{jr}g_{ri}v(z_i,z_j,z_r;\gamma) \tag{6}$$

Proof. Consider the network $g=(g_{ij}=1,g_{-ij})$, where g_{ij} is the entry at row i and column j of g; and g_{-ij} is the network g excluding entry g_{ij}. Let $g'=(g_{ij}=0,g_{-ij})$ be a network in which link ij is deleted. It is straighforward to show that for any pair i and j

$$\begin{aligned}Q(g,x,z;\theta)-Q(g',x,z;\theta) &= U_i(g,x,z;\theta)+U_j(g,x,z;\theta)-[U_i(g',x,z;\theta)+U_j(g',x,z;\theta)]\\ &= u(x_i,x_j,z_i,z_j;\alpha,\beta)+u(x_j,x_i,z_j,z_i;\alpha,\beta)\\ &+ \sum_{r\neq i,j}^{n}g_{jr}g_{ri}v(z_i,z_j,z_r;\gamma)+\sum_{r\neq i,j}^{n}g_{ir}g_{rj}v(z_j,z_i,z_r;\gamma)\end{aligned}$$

where I used the fact that $g_{ij}=g_{ji}$ in an undirected network and by assumption 2 we have $u(x_i,x_j,z_i,z_j;\alpha,\beta)=u(x_j,x_i,z_j,z_i;\alpha,\beta)$. □

The potential function is an aggregate function of the network and payoffs, that summarizes the incentives of all the players, net of the logistic matching shock. The crucial property of the potential is as follows:

$$Q(g,x,z;\theta)-Q(g',x,z;\theta)=U_i(g,x,z;\theta)+U_j(g,x,z;\theta)-[U_i(g',x,z;\theta)+U_j(g',x,z;\theta)] \tag{7}$$

where g is a network where i and j have a link, that is $g_{ij}=1$; and g' is the same network g, excluding the link between i and j, that is $g'_{ij}=0$ and $g'_{-ij}=g_{-ij}$.

The potential simplify the search for equilibrium networks. Indeed, if there are no stochastic matching shocks, the profitable deviations of i and j can be computed by using the difference in utility (the right-hand side of equation (7)) or equivalently the difference in potential (the left-hand side of equation (7)). As a consequence, it can be shown that all the networks that are pairwise stable with transfers can be found as local maxima of the potential function (Monderer and Shapley (1996), Mele (2017), Jackson and Watts (2001)).

The existence of a potential function is important because it guarantees existence of at least one equilibrium. An additional practical advantage is that one can simulate the network formation process without keeping track of each player profitable deviations: all that information is already incorporated in the potental function, which is a scalar.[35]

The network formation model is a finite state space Markov chain, because the number of networks is finite. I show that the chain is irreducible and aperiodic, therefore converging to a unique stationary distribution in the long run.

THEOREM 1. *Under Assumptions 1-3 and conditional on the community structure z, the sequence of networks generated by the model is a Markov chain with unique stationary distribution $\pi(g,x,z;\theta)$:*

$$\pi(g,x,z;\theta)=\frac{\exp\left[Q(g,x,z;\theta)\right]}{c(\theta,x,z)} \tag{8}$$

[35]The use of potential games and potential functions is common in computer science and physics. In economics, the class of congestion games is the classical example of potential games.

Proof. The proof is analogous and follows the same steps of Theorem 1 in Mele (2017), therefore it is omitted for brevity. □

The results in Proposition 1 and Theorem 1 show that the potential game characterization and the stationary distribution in equilibrium of Mele (2017), survive the additional assumptions introduced here to generate sparse networks.[36]

In the long run, the Markov chain of networks will spend most time in network configurations that have high potential. The result in Proposition 1 shows that the local maxima of the potential are pairwise stable (with transfers); therefore in the long run the networks that are most likely to be observed are pairwise stable.

I assume that the network in the data is a draw from the stationary equilibrium of the model, therefore the distribution (8) is the likelihood of observing a particular network, conditional on the community structure. The following Theorem 2 shows that the networks generated in equilibrium are sparse, provided that n is large enough.

THEOREM 2. *The networks generated by likelihood (8) are sparse. That is, there exist scalars $A > 0$ and $\lambda > 0$ such that for every pair of players i and j that belong to different communities, $z_i \neq z_j$, we have $\mathbf{E}(g_{ij}) \leq An^{-\lambda}$.*

Proof. The result follows from Lemma 3 in Appendix. □

Theorem 2 shows that as the network grows large, the unconditional probability of a link among individuals in different communities is vanishingly small. This implies that for a fixed network size n, the resulting network will display few connections across communities. However, this notion of sparsity does not impose any restriction on the density of the network within each community. In the next section I will exploit Theorem 2 to derive the asymptotic behavior of the sufficient statistics of the model.

The potential function $Q(g, x, z, \theta)$ can be decomposed in *within-* and *between-*community potentials. Let $g_{k,l}$ denote the subnetwork formed by individuals of communities $\mathcal{C}_k$ and $\mathcal{C}_l$. Let $x^{(k)}$ denote the covariates of players in community $\mathcal{C}_k$. The potential can be decomposed into the sum of sub-potentials for the sub-networks $g_{k,l}$'s and the likelihood can be written as a factorized distribution. Lemma 4 in Appendix, proves that the potential $Q(g, x, z, \theta)$ can be decomposed as the sum of subpotentials $Q_{k,l}(g_{k,l}, x^{(k)}, x^{(l)}, z; \theta)$, separating the within-community and between-community contributions as follows:

$$Q(g, x, z; \theta) = \sum_{k=1}^{K} Q_{k,k}(g_{k,l}, x^{(k)}, z; \theta) + \sum_{k=1}^{K} \sum_{l>k}^{K} Q_{k,l}(g_{k,l}, x^{(k)}, x^{(l)}, z; \theta) \tag{9}$$

$$= \sum_{k=1}^{K} \left[\sum_{i \in \mathcal{C}_k} \sum_{j \in \mathcal{C}_k} g_{ij} u(x_i, x_j, z_i, z_j; \alpha, \beta) + \frac{\gamma_k}{6} \sum_{i \in \mathcal{C}_k} \sum_{j \in \mathcal{C}_k} \sum_{r \in \mathcal{C}_k} g_{ij} g_{jr} g_{ri} \right] \tag{10}$$

$$+ \sum_{k=1}^{K} \sum_{l>k}^{K} \left[\sum_{i \in \mathcal{C}_k} \sum_{j \in \mathcal{C}_l} g_{ij} u(x_i, x_j, z_i, z_j; \alpha, \beta) \right] \tag{11}$$

[36] Additional modifications to this network formation protocol are introduced in Badev (2013), showing that the potential game characterization and the stationary equilibrium do not change.

The main consequence of this decomposition is that I can write the likelihood in factorized form

$$\pi(g,x,z;\theta)=\prod_{k=1}^{K}\frac{\exp\left[Q_{k,k}(g_{k,k},x^{(k)},z;\theta)\right]}{c_{k,k}(\mathcal{G}_{k,k},x^{(k)};\theta)}\left[\prod_{l>k}^{K}\frac{\exp\left[Q_{k,l}(g_{k,l},x^{(k)},x^{(l)},z;\theta)\right]}{c_{k,l}(\mathcal{G}_{k,l},x^{(k)},x^{(l)};\theta)}\right] \tag{12}$$

where the within-community normalizing constant is

$$c_{k,k}(\mathcal{G}_{k,k},x^{(k)};\theta)=\sum_{\omega_{k,k}\in\mathcal{G}_{k,k}}\exp\left[Q_{k,k}(\omega_{k,k},x^{(k)},z;\theta)\right] \tag{13}$$

Notice that since $Q_{k,l}(g_{k,l},x^{(k)},x^{(l)},z;\theta)=\sum_{i\in\mathcal{C}_k}\sum_{j\in\mathcal{C}_l}g_{ij}u(x_i,x_j,z_i,z_j;\alpha,\beta)$ the second part of the likelihood (12) can be written as the product of Bernoulli links,

$$\prod_{l>k}^{K}\frac{\exp\left[Q_{k,l}(g_{k,l},x^{(k)},x^{(l)},z;\theta)\right]}{c_{k,l}(\mathcal{G}_{k,l},x^{(k)},x^{(l)};\theta)}=\prod_{l>k}^{K}\prod_{i\in\mathcal{C}_k}\prod_{j\in\mathcal{C}_l}\frac{\exp\left[2g_{ij}u(x_i,x_j,z_i,z_j;\alpha,\beta)\right]}{1+\exp\left[2u(x_i,x_j,z_i,z_j;\alpha,\beta)\right]} \tag{14}$$

To summarize, the model generates independence for the links created between-communities, while the links generated within-community have strong dependence. The final result is a model that mantains the complex dependence structure of the exponential family random graphs locally, while allowing for weak dependence among links globally. This model is compatible with the exponential random graphs with local dependence developed in Schweinberger and Handcock (2015).

3. Estimation

3.1. **Model specification.** The number of parameters to estimate depends on the number of communities. Since the number of communities is not known a priori, one can assume that it could be as large as the number of players. To mitigate the problem, I adopt a parsimonious specification of the payoffs.

The utility from direct links is parameterized as

$$u(x_i,x_j,z_i,z_j;\alpha,\beta)=\alpha_{z_iz_j}+\sum_{p=1}^{P}\beta_p f_p(x_i,x_j) \tag{15}$$

The first part of the utility is parameter $\alpha_{z_iz_j}$, which models the cost of forming a link, and it can vary with the communities. The second part is the benefit of forming a link, that I assume is a function of covariates. Example of possible functions f_p's are:

$$\begin{aligned} f_p(x_i,x_j) &= |x_i-x_j|; \qquad & f_p(x_i,x_j)&=\mathbf{1}_{\{x_i=x_j\}}\\ f_p(x_i,x_j) &= \mathbf{1}_{\{x_i=x_j=a\}}; \qquad & f_p(x_i,x_j)&=x_i+x_j\\ f_p(x_i,x_j) &= x_i\cdot x_j \end{aligned}$$

Notice that if there are K communities, the vector α consists of at least $K(K-1)/2$ parameters, while γ is a vector of K parameters; the length of β depends on the number P of functions f_p, which will eventually depend on the specification chosen by the researcher and the number of columns of the covariate matrix x. Therefore, our model has at least

$K(K+1)/2+P$ parameters to estimate. Since the number of communities is not known, we may potentially have as many as n communities, thus $n(n+1)/2+P$ parameters.

To keep the model parsimonious, I constrain the cost of forming links across communities to be the same for each player. On the other hand, the cost of forming links to people in the same community is allowed to vary by community.[37]

$$\alpha_{z_i z_j} = \begin{cases} \alpha_k & \text{if } z_i = z_j \text{ and } z_{ik} = 1, \text{ for } k = 1, 2, ..., K \\ \alpha_b & \text{otherwise} \end{cases} \tag{16}$$

If players i and j belong to the same community $\mathcal{C}_k$, their cost of linking is α_k. Otherwise, the cost of forming a link is α_b.

I include the following covariates in the models: race, gender, grade and parental income. The final specification of the utility function is

$$\begin{aligned} U_i(g,x,z;\theta) \;=\; & \sum_{j=1}^{n} g_{ij}\big[\alpha_{z_i z_j} \\ & + \beta_{white,white}\mathbf{1}_{\{race_i=race_j=white\}} + \beta_{black,black}\mathbf{1}_{\{race_i=race_j=black\}} \\ & + \beta_{hisp,hisp}\mathbf{1}_{\{race_i=race_j=hispanic\}} + \beta_{grade7,grade7}\mathbf{1}_{\{grade_i=grade_j=7\}} \\ & + \beta_{grade8,grade8}\mathbf{1}_{\{grade_i=grade_j=8\}} + \beta_{grade9,grade9}\mathbf{1}_{\{grade_i=grade_j=9\}} \\ & + \beta_{grade10,grade10}\mathbf{1}_{\{grade_i=grade_j=10\}} + \beta_{grade11,grade11}\mathbf{1}_{\{grade_i=grade_j=11\}} \\ & + \beta_{grade12,grade12}\mathbf{1}_{\{grade_i=grade_j=12\}} + \beta_{male,male}\mathbf{1}_{\{gender_i=gender_j=male\}} \\ & + \beta_{female,female}\mathbf{1}_{\{gender_i=gender_j=female\}} \\ & + \beta_{|income_i-income_j|}|income_i - income_j| \\ & + \sum_{r} g_{jr}g_{rj}\gamma(z_i,z_j,z_r)\big] \end{aligned} \tag{17}$$

Our specification allows for homophily in race, gender, grade and (parental) income. Homophily in unobservables is captured by the difference between the cost parameters $\alpha_{z_i z_j}$ estimated within community and across communities (α_b). The clustering is captured by the parameter $\gamma(z_i, z_j, z_r)$, related to the preferences for common friends.

3.2. **Asymptotic normality of sufficient statistics.** The model in this paper is an exponential family with normalizing constant. Given the utility function specification in (17), the potential function is linear and it can be written in the form

$$Q(g,x,z;\theta) = \sum_{p=1}^{P} \theta_p S_p(g,x,z) \tag{18}$$

for $p = 1, ..., P$. The scalars θ_p's are parameters, while the functions $S_p(g, x, z)$ are sufficient statistics of the model. For example, the sufficient statistics relative to the cost of linking is

$$S_1(g,x,z) := \sum_{i=1}^{n}\sum_{j=1}^{n} g_{ij} \tag{19}$$

[37] This is also the advice provided in Schweinberger and Handcock (2015) for their hierarchical model.

while the sufficient statistics for homophily of white students is

$$S_2(g,x,z) := \sum_{i=1}^{n}\sum_{j=1}^{n} g_{ij}\mathbf{1}_{\{race_i=race_j=white\}} \tag{20}$$

The other sufficient statistics relative to the remaining homophily terms are derived analogously. Finally, the sufficient statistics for clustering is

$$S_{14}(g,x,z) := \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{r=1}^{n} g_{ij}g_{jr}g_{ri} \tag{21}$$

The following theorem shows that the sufficient statistics are asymptotically normal, as the number of communities K grows large. This is a general result and it holds for more general specifications than the one in equation (17).[38]

THEOREM 3. *If the meeting technology parameter $\delta > 3$, then the sufficient statistics of the network formation model are asymptotically normal. That is, as the number of communities K grows large, the p-th sufficient statistics $S_p(g,x,z)$, normalized by its standard deviation, converges in distribution to a normal random variable*

$$\frac{S_p(g,x,z)}{\sqrt{V\left[S_p(g,x,z)\right]}} \xrightarrow{d} N(0,1) \text{ as } K\to\infty \tag{22}$$

where $V\left[S_p(g,x,z)\right]$ is the variance of sufficient statistics $S_p(g,x,z)$.

Proof. The proof is contained in Appendix B. □

Theorem 3 shows that the model is well-behaved: most of the probability mass is placed around the expected value of the sufficient statistics. The two most important ingredients that lead to the result in Theorem 3 are the sparsity described in Theorem 2 and the ability to factorize the likelihood into within- and between-communities components; these properties imply weak dependence among links that we can exploit to obtain the asymptotic result.

3.3. **Estimation strategy.** The main challenge in estimation of the model is that the community structure is unknown. The Bayesian approach is to impose a prior on the number of communities and use that to estimate the posterior. I follow Schweinberger and Handcock (2015) and use their nonparametric priors to model the communities. The advantage of this approach is that the prior can be truncated at a maximum of $K_{\max}$ communities. More details about priors and prior truncation are provided in the Appendix.

The likelihood of the model can be written as

$$L\left(g,Z;\theta,\eta,x\right) = \sum_{z\in\mathcal{Z}} P_\theta\left(G=g|X=x,Z=z\right)P_\eta\left(Z=z\right) \tag{23}$$

[38]The details are provided in Appendix B.

The first part is the likelihood of observing network g in the long run, given the covariates x and the community structure z, therefore

(24)
$$P_\theta\left(G=g|X=x,Z=z\right)=\prod_{k=1}^{K}\frac{\exp\left[Q_{k,k}(g_{k,k},x^{(k)},z)\right]}{c_{k,k}(\mathcal{G}_{k,k},x^{(k)};\theta)}\left[\prod_{l>k}^{K}\frac{\exp\left[Q_{k,l}(g_{k,l},x^{(k)},x^{(l)},z)\right]}{c_{k,l}(\mathcal{G}_{k,l},x^{(k)},x^{(l)};\theta)}\right]$$

For the community structure, I follow a standard assumption in the stochastic blockmodels literature (Airoldi et al. (2008), Schweinberger and Handcock (2015)), and use a multinomial distribution. The community assignment to each player is i.i.d.

$$Z_i|\eta_1,...,\eta_K \overset{iid}{\sim} Multinomial\left(1;\eta_1,...\eta_K\right) \text{ for } i=1,...,n \tag{25}$$

Following Ishwaran and James (2001) and Schweinberger and Handcock (2015), I use the following nonparametric prior for η_k, $k=1,2,3,...,K$,

$$\eta_1 = V_1 \tag{26}$$

$$\eta_k = V_k\prod_{j=1}^{k-1}(1-V_j) \qquad k=2,3,4,... \tag{27}$$

$$V_k|\phi \overset{iid}{\sim} Beta(1,\phi) \qquad k=1,2,3,... \tag{28}$$

$$\phi>0 \quad \text{and} \quad \sum_{k=1}^{\infty}\eta_k=1 \ w.p.1 \tag{29}$$

As a practical matter, the estimation of the posterior is computationally expensive, especially when the number of communities K is large. In principle, one can have as many as n communities. For moderate size networks with $n<200$ this would mean the estimation of a model with at least 400 parameters. Therefore, I truncate the prior for the communities. The details are shown in appendix.

The priors for the payoffs are multivariate normals

$$\alpha_b|\mu_b,\Sigma_b \sim MVN\left(\mu_b,\Sigma_b\right) \tag{30}$$

$$(\alpha_k,\gamma_k)|\mu_w,\Sigma_w \sim MVN\left(\mu_w,\Sigma_w\right) \text{ for } k=1,..,K_{max} \tag{31}$$

$$\beta|\mu_\beta,\Sigma_\beta \sim MVN\left(\mu_\beta,\Sigma_\beta\right) \tag{32}$$

I adopt a hierarchical approach and speficy hyper-priors for each parameter. For details see Appendix.

3.4. **Sampling from the posterior.** The complex form of the likelihood does not allow direct sampling from the posterior. I rely on the exchange MCMC method developed in Murray et al. (2006) and Liang (2010).[39] The posterior distribution can be written as follows

$$\begin{aligned} p(\phi,\mu_w,\Sigma_w,\mu_b,\Sigma_b,\mu_\beta,\Sigma_\beta,\eta,\alpha,\beta,\gamma,z|g,x) &\propto p(\phi,\mu_w,\Sigma_w,\mu_b,\Sigma_b,\mu_\beta,\Sigma_\beta,\eta,\alpha,\beta,\gamma) \\ &\times P_\eta\left(Z=z\right)P_\theta\left(G=g|X=x,Z=z\right) \end{aligned} \tag{33}$$

[39]The application of the exchange algorithm to network models has been shown in Caimo and Friel (2011), Atchade and Wang (2014) and Mele (2017).

where $p(\phi, \mu_w, \Sigma_w, \mu_b, \Sigma_b, \mu_\beta, \Sigma_\beta, \eta, \alpha, \beta, \gamma)$ is the prior distribution. The prior is assumed to factorize in the following form

$$
\begin{aligned}
(34) \quad p(\phi, \mu_w, \Sigma_w, \mu_b, \Sigma_b, \mu_\beta, \Sigma_\beta, \eta, \alpha, \beta, \gamma) &= p(\phi)p(\mu_w)p(\Sigma_w)p(\mu_b)p(\Sigma_b)p(\mu_\beta)p(\Sigma_\beta) \\
&\times \; p(\eta|\phi)p(\alpha_b|\mu_b, \Sigma_b)p(\beta|\mu_\beta, \Sigma_\beta) \\
&\times \; \left[\prod_{k=1}^{K_{max}} p(\alpha_k, \gamma_k|\mu_w, \Sigma_w)\right]
\end{aligned}
$$

The details of the sampler are provided in Murray et al. (2006), Mele (2017), Caimo and Friel (2011), Liang (2010) and Schweinberger and Handcock (2015).[40]

The exchange algorithm for this model is slightly different from the original Murray et al. (2006) and Liang (2010)'s sampler used in Mele (2017). The main additional complication is that the communities are unknown, and therefore need to be treated as an additional parameter in the sampler. The algorithm augments the posterior parameters with auxiliary variables g^*, z^* and $\theta^* := (\alpha^*, \beta^*, \gamma^*)$, proceeding with the following steps at each iteration:

(1) Sample (θ^*, z^*) from auxiliary distribution $q(\theta^*, z^*|\eta, \theta, z, g)$
(2) Sample g^* from $\pi(\omega, x, z^*; \theta^*)$ using the Metropolis-Hastings sampler[41]
(3) Propose to swap (θ, z) with (θ^*, z^*), accepting with probability $\min\{1, exch\}$, where $exch$ is

$$
\begin{aligned}
(35) \qquad exch = &\frac{P_\eta(Z = z^*)}{P_\eta(Z = z)} \frac{q(\theta, z|\eta, \theta^*, z^*, g)}{q(\theta^*, z^*|\eta, \theta, z, g)} \frac{\pi(g, x, z^*; \theta^*)}{\pi(g, x, z; \theta)} \frac{\pi(g^*, x, z; \theta)}{\pi(g^*, x, z^*; \theta^*)} \\
&\times \frac{\prod_{k=1}^{K_{max}} p(\alpha_k^*, \gamma_k^*|\mu_w, \Sigma_w)}{\prod_{k=1}^{K_{max}} p(\alpha_k, \gamma_k|\mu_w, \Sigma_w)}
\end{aligned}
$$

and $P_\eta(Z = z)$ is the multinomial distribution that generates the community structure, $\pi(g, x, z; \theta)$ is the stationary distribution of the model conditional on community structure, and $p(\alpha_k, \gamma_k|\mu_w, \Sigma_w)$ are the priors.

The practical implication of the formula for acceptance probability $exch$, is that the normalizing constants included in the discrete exponential distribution cancel out. Therefore the sampling using the exchange algorith is feasible, while sampling from the posterior using standard Metropolis or Gibbs sampler is infeasible. For a formal discussion see Mele (2017), Appendix B.

The auxiliary distribution $q(\theta^*, z^*|\eta, \theta, z, g)$ proposes θ^* that are Gaussians centered at θ and z^* that are generated from the full conditional distribution of the community memberships.[42] The reason for such updates is that these local moves do not lead to a very high rejection rate of the exchange algorithm, as pointed out in Caimo and Friel (2011) and Mele (2017).

[40] I use the package `hergm` in `R`, developed by Schweinberger and Luna (forthcoming), to estimate the model. All the codes for replication are available from the author.

[41] This is the same algorithm for network simulation used in Mele (2017).

[42] Additional details and the full set of updates are in the Supplement of Schweinberger and Handcock (2015) and in Schweinberger and Luna (forthcoming). The package `hergm` in `R` implements these methods.

3.5. **Identification and label invariance.** An additional challenge is that the likelihood of this model is invariant to permutations of the community labels. This problem is common in the literature on finite mixture models, where the likelihood is invariant to permutations of the labels of the mixture's components (Gelman et al. (2003), McLachlan and Peel (2000), Stephens (2000)). This complicates inference for the community-specific parameters, because the community labels may switch several times during the MCMC simulation.

Nonetheless, the use of nonparametric priors implies that the full posterior is not invariant to permutations of the community labels. This reduces the problem. Furthermore, after obtaining a MCMC sample from the posterior distribution, I use the algorithm of Schweinberger and Handcock (2015) to relabel the output of the posterior simulations. This approach is common in the Bayesian literature on finite mixture models.

Suppose to have a MCMC posterior simulation $\{\theta^s, z^s\}_{s=1}^n$ of length S. The relabeling algorithm minimizes the loss function

$$L(\xi, \nu(Z)) = \min_{\nu} L_0 [\xi, \nu(Z)] \tag{36}$$

where

$$L_0 [\xi, \nu(Z)] = -\log \prod_{i=1}^{n} \xi_{i,\mathcal{C}_i} \tag{37}$$

where ξ is an $n \times K$ matrix whose entry $\xi_{i,k}$ is the probability that individual i is reported to be in community/type k; and $\nu(Z)$ is a permutation of the community structure Z.

So the goal of the relabeling algorithm is to choose the matrix ξ that minimizes the posterior expectation of loss function $L[\xi, \nu(Z)]$. In practice the posterior expectation is approximated by the Monte Carlo sample

$$\frac{1}{S}\sum_{s=1}^{S} \min_{\nu_s} [L_0 [\xi, \nu_s(z^s)]] = \min_{\nu_1,...,\nu_S} \left[\frac{1}{S}\sum_{s=1}^{S} L_0 [\xi, \nu_s(z^s)]\right] \tag{38}$$

and the algorithm starts from some initial permutation of the community labels $\nu_1, \nu_2, ..., \nu_S$ and iterates on the following two steps until convergence:

(1) choose $\hat{\xi}$ to minimize $\sum_{s=1}^{S} [L_0 [\xi, \nu_s(z^s)]]$ subject to the constraint $\sum_{k=1}^{K_{max}} \xi_{i,k} = 1$ for $i = 1, ..., n$;
(2) for $s = 1, ..., S$ choose ν_s to minimize $L_0 [\xi, \nu_s(z^s)]$

The second step is infeasible unless the number of communities K_{max} is very small. Schweinberger and Handcock (2015)'s implementation uses Simulated Annealing to perform the S minimizations in parallel, leading to a more practical algorithm.[43]

3.6. **Data.** The network used in the estimation exercise is from the *National Longitudinal Study of Adolescent Health* (Add Health). This dataset contains information on a nationally

[43]Similar algorithms for relabeling the output of the MCMC are discussed in Gelman et al. (2003) and McLachlan and Peel (2000). More details on the practical implementation are in Schweinberger and Handcock (2015) and Stephens (2000).

representative sample of US schools. The survey started in 1994, when the 90118 participants were entering grades 7-12, and the project collected data in four successive waves.[44] Each student responded to an *in-school* questionnaire, and a subsample of 20745 was given an *in-home* interview to collect more detailed information about behaviors, characteristics and health status. The survey asked each student a long set of demographic, health and socioeconomic questions. In addition, students were provided with the roster of their school and asked to identify up to 5 male and 5 female friends.[45] I use this part of the survey to construct the (undirected) network of friendships.

The model estimated includes the following covariates: race, gender, grade and parental income. These are some of the variables that are considered good predictors of friendships during adolescence.[46]

In this paper I use only data from school 28 and Wave I (1994), from the saturated sample. Each student in this sample completed both the in-school and in-home questionnaires, and the researchers made a significant effort to avoid any missing information on the students.

I use data on racial group, grade and gender of individuals. A student with a missing value in any of these variables is dropped from the sample. Each student that declares to be of Hispanic origin is considered Hispanic. The remaining non-Hispanic students are assigned to the racial group they declared. Therefore the racial categories are: White, Black, Asian, Hispanic and Other race. Other race contains Native Americans. I also control for homophily in income, using the family income reported in a question from the parent questionnaire.[47]

The school contains 150 students, with 58.7% females. The school is very racially heterogeneous: 42% Whites/Caucasians, 45.3% African-Americans, 0.7% Asians, 10.7% Hispanics and 1.3% Other race. The racial fragmentation index is 0.606. The school offers all grades from 7 to 12, with a relatively balanced population among the different age groups, respectively 17.3%, 17.3%, 20%, 16.7%, 14%, and 14.7%.

This school exhibits a high level of segregation, measure using the index developed by Freeman (1972), that varies between a minimum of 0 (no segregation) and 1 (perfect segregation). The measured segregation level is 0.72 for Whites/Caucasians, 0.764 for African Americans, and 0.429 for Hispanics. The segregation by gender is 0.255.

4. Empirical Results

4.1. **Number of communities.** There is no a priori way to select the model and/or determine the number of communities for the prior distribution. One can certainly impose a prior on K, but this has a computational cost. I consider a more heuristic approach,

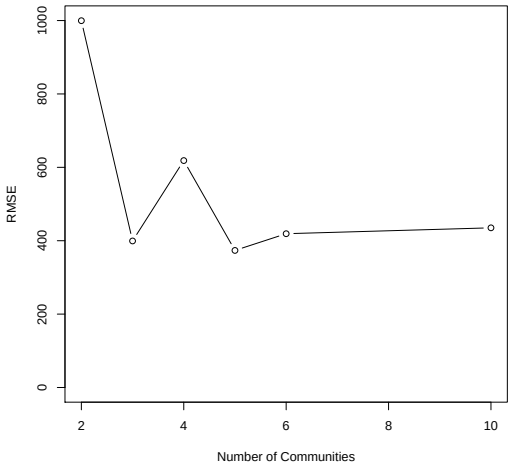
[44]More details about the sampling design and the representativeness are contained in Moody (2001) and the Add Health website http://www.cpc.unc.edu/projects/addhealth/projects/addhealth

[45]One can think that this limit could bias the friendship data, but only 3% of the students nominated 10 friends (Moody, 2001). Moreover, the estimation routine could be easily extended to deal with missing links.

[46]See for example, Moody (2001), Mayer and Puller (2008), Boucher (2015).

[47]There are several cases in which the family income is missing. For those observations, I imputed values drawn from the unconditional income distribution of the community. An alternative but computationally very costly alternative is to introduce an additional step in the simulation, in which the imputation of missing incomes is done at each iteration.

FIGURE 1. Posterior predictions of root mean squared error for the number of triangles in the network



following Schweinberger and Handcock (2015) and the goodness-of-fit procedures from the ERGM literature (Snijders (2002), Koskinen (2008)). I choose the number of communities by trying different K's in increasing order, $K = \{2, 3, 4, 5, ...\}$ and checking whether the estimated model is able to replicate some empirical properties of the network in the data, that is the number of links and triangles. This procedure delivers the most parsimonious model that is able to satisfy the empirical properties of the observed network. The chosen K_{max} is the smallest K for which the root mean squared error for triangles in the network does not improve anymore. Figure 1 shows the root mean squared error for the posterior prediction of the number of triangles in the network, for several models. The main results presented here are the estimates from a model with $K_{max} = 3$, since there is no much improvement in the RMSE when increasing the number of communities to $K_{max} = 4$. I report the estimated community membership probabilities estimated with $K_{max} = \{4, 5, 6, 10\}$ in Figure 7 in Appendix. Structural parameter estimates for $K_{max} = 5$, are reported in Table 2 in Appendix, for completeness.[48]

4.2. **Estimated structural parameters.** The estimated structural parameters are shown in Table 1, where I report the specification with $K_{max} = 3$. I include mean, standard deviation, median and 2.5% and 97.5% quantiles of the posterior distribution. The histograms of the marginal posteriors are in Figure 4, 5 and 6 in Appendix.

Most parameters are estimated precisely, with the exception of γ_3, the payoff from common friends in the third community; and the parameters relative to homophily by gender ($\beta_{male,male}$ and $\beta_{female,female}$).

In panel A, I show the marginal posterior for the cost. As expected, the estimated α's are negative; the cost of forming links across communities (α_b) is higher than the cost of forming links within communities ($\alpha_1, \alpha_2, \alpha_3$). While the cost of forming links within communities

[48]The estimates for all the values of K_{max} are available from the author.

does not differ significantly for students of type 1 and 2, individuals in community 3 seem relatively more social.

Panel B reports the estimates for terms including covariates. There is homophily by race, as the marginal utility of a link increases when players form a link with a student of the same racial group. The same effect is present for grade. The estimates for gender are close to zero and very widely spread. I also estimate homophily by income, as the coefficient $\beta_{|income_i - income_j|}$ is negative.

Panel C shows the estimates for payoffs from common friends. An additional common friend is more valuable for students of type 2 than type 1. For student of type 3 the estimate is not very precise, but it is nonetheless positive on average.

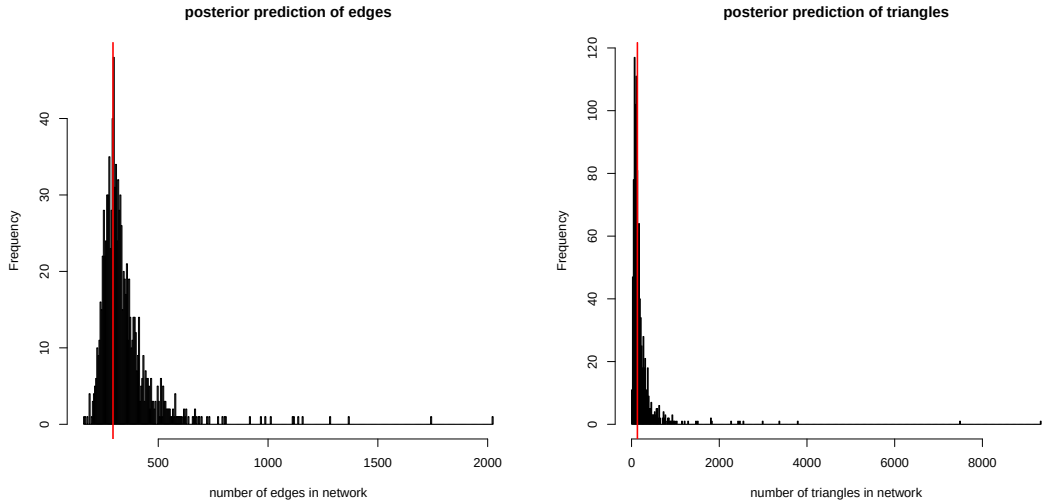
TABLE 1. Estimated posterior of the structural parameters ($K_{max} = 3$)

Parameter	**Post. mean**	**Post. s.d.**	**Posterior quantiles** 2.5%	50%	97.5%
A. Cost of link					
α_1	-4.070	0.464	-4.888	-4.086	-3.091
α_2	-3.854	0.587	-4.883	-3.895	-2.589
α_3	-2.527	1.049	-4.385	-2.609	-0.316
α_b	-5.754	0.455	-6.636	-5.763	-4.837
B. Payoff from covariates					
$\beta_{white,white}$	1.002	0.246	0.500	1.017	1.420
$\beta_{black,black}$	0.923	0.252	0.424	0.938	1.364
$\beta_{hisp,hisp}$	1.965	0.628	0.789	1.920	3.128
$\beta_{grade7,grade7}$	1.371	0.290	0.685	1.409	1.831
$\beta_{grade8,grade8}$	1.321	0.311	0.627	1.327	1.892
$\beta_{grade9,grade9}$	1.203	0.332	0.568	1.172	1.883
$\beta_{grade10,grade10}$	1.140	0.446	0.207	1.127	1.929
$\beta_{grade11,grade11}$	1.241	0.433	0.249	1.291	1.973
$\beta_{grade12,grade12}$	1.029	0.281	0.435	1.033	1.562
$\beta_{male,male}$	-0.061	0.297	-0.689	-0.029	0.450
$\beta_{female,female}$	-0.170	0.254	-0.725	-0.135	0.294
$\beta_{\lvert income_i - income_j \rvert}$	-0.588	0.278	-1.208	-0.568	-0.136
C. Payoff from common friends					
γ_1	0.969	0.149	0.644	0.977	1.244
γ_2	1.573	0.562	0.508	1.561	2.738
γ_3	0.995	0.948	-0.889	0.969	2.920

The estimates are obtained from a run of 100,000 steps of the exchange algorithm, collecting a posterior sample of 8000 draws. Panel A shows the estimates of linking costs; Panel B shows the estimates of the homophily terms; Panel C shows the estimates for the common friends' payoffs. I report mean, standard deviation, median, the 2.5% and 97.5% quantiles.

I conclude that there is strong homophily by observable characteristics, especially race, grade and income. In addition, the estimated marginal posteriors show that there is some

FIGURE 2. Posterior predictions of number of links and triangles in the network



The posterior predictions are obtained by a 1000 simulations from the posterior estimated in Table 1. The red line is the value observed in the data

heterogeneity across communities in both costs of forming links and payoffs from common friends. Community 3 seems to be the most social: student in this community have the lowest costs of forming links. Students in community 2 care more about common friends.

4.3. **Fit of the model.** The model fit is relatively good: our posterior predictions are able to match the observed links, triangles and homophily. Figure 2 shows the histogram of posterior predictions for the number of links in the network and the number of triangles. I simulated 1000 realizations of the network, drawing from the posterior distribution of the parameters. The observed number of links is 294 and the posterior mean prediction is 346.5, with median prediction of 317.

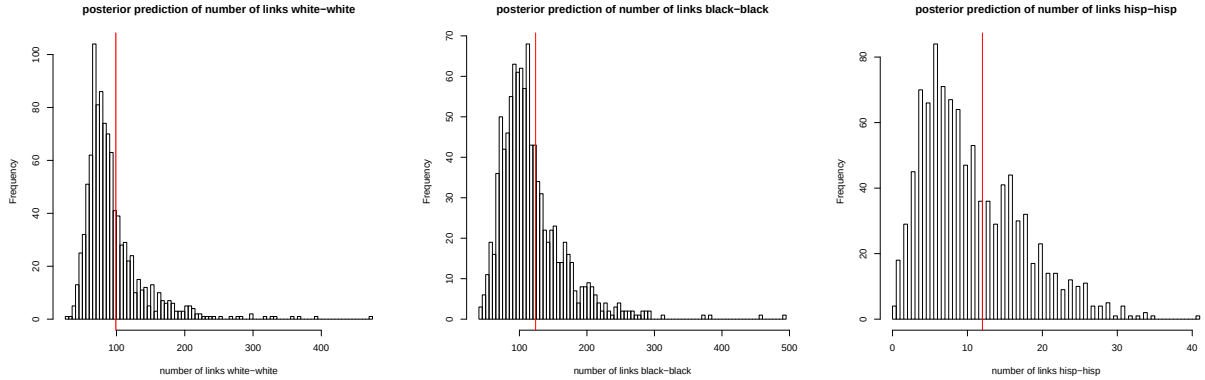
It is well known that the number of triangles, is the most difficult statistics to match. Diaconis and Chatterjee (2013) and Mele (2017) show that in exponential-family random graphs the number of triangles tends to be degenerate, either very close to zero or very close to the maximum number. In this network, there are 133 triangles and the posterior mean prediction is 225.6, with a median prediction of 133. There are some extreme values in the posterior simulations but the general fit is good.

Figure 3 shows that the model is able to replicate also the homophily by race. I report the histograms of posterior predictions for the observed number of friendships among whites, blacks and hispanics in the school. The red vertical line is the observed value. As in the figures above, while there are some extreme values, the fit of the model is very good.

The observed number of friendships among whites, african-american and hispanic students are 99, 124 and 12, respectively. The predicted posterior means are 94.9, 118.4 and 10.8; the predicted posterior medians are 83, 108 and 9 respectively.

I conclude that the model is able to replicate the empirical aggregate features of the network, both in terms of triadic closure and homophily.

FIGURE 3. Posterior predictions for racial homophily



The posterior predictions are obtained by a 1000 simulations from the posterior estimated in Table 1. The red line is the value observed in the data.

5. Conclusion

Social networks have three important empirical features: homophily, clustering and sparsity. This paper develops a structural model of network formation that generates homophily and clustering in equilibrium, generating sparse networks. Players belongs to different communities that affect their cost of linking and their probability of metting other individuals. Agents care about the composition of their links (homophily) and the number of common friends (clustering). The probability of meeting people in a different community decreases with the size of the network. These assumptions generate a sequence of meetings and link update that in the long-run converges a discrete exponential family random graph.

I show that the model is able to match the empirical features of Add Health school friendship data, while also providing insights on the economics of network formation. People in different communities have different costs of linking and have different payoffs from common friends, suggesting that unobserved characteristics may affect network formation as much as observables. Therefore it is important to develop tractable models of empirical network formation that can deal with unobserved heterogeneity.

References

Acemoglu, D., M. Dahleh, I. Lobel and A. Ozdaglar (2011), 'Bayesian learning in social networks', *Review of Economic Studies* **78**(4), 1201–1236.

Airoldi, Edoardo M., David Blei, Stephen E. Fienberg and Eric P. Xing (2008), 'Mixed membership stochastic blockmodels', *Journal of Machine Learning* **9**, 1981–2014.

Aristoff, David and Lingjiong Zhu (2014), On the phase transition curve in a directed exponential random graph model.

Atchade, Yves and Jing Wang (2014), 'Bayesian inference of exponential random graph models for large social networks', *Communications in Statistics - Simulation and Computation* **43**(2), 359–377.

Badev, Anton (2013), Discrete games in endogenous networks: Theory and policy.

Bala, Venkatesh and Sanjeev Goyal (2000), 'A noncooperative model of network formation', *Econometrica* **68**(5), 1181–1229.

Berry, Steve, Ahmed Khwaja, Vineet Kumar, Andres Musalem, Kenneth C. Wilbur, Greg Allenby, Bharat Anand, Pradeep Chintagunta, W. Michael Hanemann, Przemek Jeziorski and Angelo Mele (2014), 'Structural models of complementary choices', *Marketing Letters* **25**(3), 245–256.

Blume, Lawrence E. (1993), 'The statistical mechanics of strategic interaction', *Games and Economic Behavior* **5**(3), 387–424.

Bonhomme, Stephane, Thibaut Lamadon and Elena Manresa (2017), Discretizing unobserved heterogeneity. Working Paper.

Boucher, Vincent (2015), 'Structural homophily', *The International Economic Review* **56**(1), 235–264.

Boucher, Vincent and Ismael Mourifie (forthcoming), 'My friend far far away: A random field approach to exponential random graph models', *Econometrics Journal* .

Breza, Emily, Arun Chandrasekhar, Tyler Mccormik and Mengjie Pan (2017), Using aggregated relational data to feasibly identify network structure without network data.

Butts, Carter (2009), Using potential games to parameterize erg models. working paper.

Caimo, Alberto and Nial Friel (2011), 'Bayesian inference for exponential random graph models', *Social Networks* **33**(1), 41–55.

Calvo-Armengol, Antoni, Eleonora Patacchini and Yves Zenou (2009), 'Peer effects and social networks in education', *Review of Economic Studies* **76**, 1239–1267.

Chandrasekhar, Arun G. (2016), *Oxford handbook on the economics of networks.*, Oxford Univerisity Press, chapter Econometrics of network formation.

Chandrasekhar, Arun and Matthew Jackson (2014), Tractable and consistent exponential random graph models.

Chandrasekhar, Arun and Matthew Jackson (2016), A network formation model based on subgraphs. working paper.

Charbonneau, Karyne B. (2017), 'Multiple fixed effects in binary response panel data models', *Econometrics Journal* **20**(3), S1–S13.

Chatterjee, Sourav and S.R.S. Varadhan (2011), 'The large deviation principle for the erdos-renyi random graph', *European Journal of Combinatorics* **32**(7), 1000 – 1017. Homomorphisms and Limits.

Christakis, Nicholas, James Fowler, Guido W. Imbens and Karthik Kalyanaraman (2010), An empirical model for strategic network formation. Harvard University.

Conley, Timothy and Christopher Udry (2010), 'Learning about a new technology: Pineapple in ghana', *American Economic Review* **100**(1), 35–69.

Conley, Timothy G. (1999), 'Gmm estimation with cross sectional dependence', *Journal of Econometrics* **92**(1), 1–45.

Conley, Timothy G. and Giorgio Topa (2007), 'Estimating dynamic local interactions models', *Journal of Econometrics* **140**(1), 282–303.

Currarini, Sergio, Matthew O. Jackson and Paolo Pin (2009), 'An economic model of friendship: Homophily, minorities, and segregation', *Econometrica* **77**(4), 1003–1045.

Currarini, Sergio, Matthew O. Jackson and Paolo Pin (2010), 'Identifying the roles of race-based choice and chance in high school friendship network formation', *the Proceedings of*

the National Academy of Sciences **107**(11), 4857–4861.

De Giorgi, Giacomo, Michele Pellizzari and Silvia Redaelli (2010), 'Identification of social interactions through partially overlapping peer groups', *American Economic Journal: Applied Economics* **2**(2).

DePaula, Aureo (forthcoming), 'Econometrics of network models', *Advances in Economics and Econometrics: Theory and Applications* .

DePaula, Aureo, Seth Richards-Shubik and Elie Tamer (forthcoming), 'Identifying preferences in networks with bounded degree', *Econometrica* .

Diaconis, Persi and Sourav Chatterjee (2013), 'Estimating and understanding exponential random graph models', *Annals of Statistics* **41**(5), 2428–2461.

Dzemski, Andreas (2017), An empirical model of dyadic link formation in a network with unobserved heterogeneity. Working Paper.

Echenique, Federico and Roland Fryer (2007), 'A measure of segregation based on social interactions', *Quarterly Journal of Economics* **122**(2), 441–485.

Fafchamps, Marcel and Flore Gubert (2007), 'Risk sharing and network formation', *American Economic Review Papers and Proceedings* **97**(2), 75–79.

Fox, Jeremy and Natalia Lazzati (forthcoming), 'A note on identification of discrete choice models for bundles and binary games', *Quantitative Economics* .

Freeman, L. (1972), 'Segregation in social networks', *Sociological Methods and Research* **6**, 411–427.

Galeotti, Andrea (2006), 'One-way flow networks: the role of heterogeneity', *Economic Theory* **29**(1), 163–179.

Gelman, A., J. Carlin, H. Stern and D. Rubin (2003), *Bayesian Data Analysis, Second Edition*, Chapman & Hall/CRC.

Gentzkow, Matthew (2007), 'Valuing new goods in a model with complementarities: online newspapers', *American Economic Review* **97**(3), 713–744.

Goldsmith-Pinkham, Paul and Guido W. Imbens (2013), 'Social networks and the identification of peer effects', *Journal of Business and Economic Statistics* **31**(3), 253–264.

Golub, Benjamin and Matthew Jackson (2011), 'Network structure and the speed of learning: Measuring homophily based on its consequences', *Annals of Economics and Statistics* .

Graham, Bryan (2017), 'An empirical model of network formation: with degree heterogeneity', *Econometrica* **85**(4), 1033–1063.

Heckman, James J. (1978), 'Dummy endogenous variables in a simultaneous equation system', *Econometrica* **46**(4), 931–959.

Hsieh, Chih-Sheng and Lung-Fei Lee (2012), A structural modeling approach for network formation and social interactions with applications to students' friendship choices and selectivity on activities.

Ishwaran, Hemant and Lancelot F James (2001), 'Gibbs sampling methods for stick-breaking priors', *Journal of the American Statistical Association* **96**(453), 161–173.

Jackson, Matthew and Alison Watts (2001), 'The existence of pairwise stable networks', *Seoul Journal of Economics* **14**(3), 299–321.

Jackson, Matthew and Asher Wolinsky (1996), 'A strategic model of social and economic networks', *Journal of Economic Theory* **71**(1), 44–74.

Jackson, Matthew O. (2008), *Social and Economics Networks*, Princeton.

Jackson, Matthew O. and Brian W. Rogers (2007), 'Meeting strangers and friends of friends: How random are social networks?', *American Economic Review* **97**(3), 890–915.

Jochmans, Koen (2017), 'Two-way models for gravity', *Review of Economics and Statistics* .

Koskinen, Johan H. (2008), The linked importance sampler auxiliary variable metropolis hastings algorithm for distributions with intractable normalising constants. MelNet Social Networks Laboratory Technical Report 08-01, Department of Psychology, School of Behavioural Science, University of Melbourne, Australia.

Kosyakova, Tetyana, Sanjog Misra, Thomas Otter and Christian Neuerburg (2017), Measuring substitution and complementarity in menu based choice experiments. working paper.

König, Michael David (2016), 'The formation of networks with local spillovers and limited observability', *Theoretical Economics* **11**(3), 813–863.

Laschever, Ron (2009), The doughboys network: Social interactions and labor market outcomes of world war i veterans. working paper.

Lehman, E. L. (1983), *Theory of Point Estimation*, Wiley and Sons.

Leung, Michael (2014), A random-field approach to inference in large models of network formation. working paper.

Liang, Faming (2010), 'A double metropolis-hastings sampler for spatial models with intractable normalizing constants', *Journal of Statistical Computing and Simulation* **80**, 1007–1022.

Lovasz, L. (2012), *Large Networks and Graph Limits*, American Mathematical Society colloquium publications, American Mathematical Society.

Marjoram, Paul, John Molitor, Vincent Plagnol and Simon Tavaré (2003), 'Markov chain monte carlo without likelihoods', *Proceedings of the National Academy of Sciences* **100**(26), 15324–15328.

Mayer, Adalbert and Steven L. Puller (2008), 'The old boy (and girl) network: Social network formation on university campuses.', *Journal of Public Economics* **92**(1-2), 329–347.

McLachlan, G. J. and D. Peel (2000), *Finite mixture models*, Wiley Series in Probability and Statistics.

Mele, Angelo (2017), 'A structural model of dense network formation', *Econometrica* **85**, 825–850.

Mele, Angelo and Lingjiong Zhu (2017), Approximate variational estimation for a model of network formation.

Menzel, Konrad (2015), Strategic network formation with many agents, Working papers, NYU.

Monderer, Dov and Lloyd Shapley (1996), 'Potential games', *Games and Economic Behavior* **14**, 124–143.

Moody, James (2001), 'Race, school integration, and friendship segregation in america', *American Journal of Sociology* **103**(7), 679–716.

Murray, Iain A., Zoubin Ghahramani and David J. C. MacKay (2006), 'Mcmc for doubly-intractable distributions', *Uncertainty in Artificial Intelligence* .

Nakajima, Ryo (2007), 'Measuring peer effects on youth smoking behavior', *Review of Economic Studies* **74**(3), 897–935.

Ridder, Geert and Shuyang Sheng (2015), Estimation of large network formation games, Working papers, UCLA.

Schweinberger, Michael and Mark S Handcock (2015), 'Local dependence in random graph models: char- acterization, properties and statistical inference.', *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* (77), 1–30.

Schweinberger, Michael and Pamela Luna (forthcoming), 'Hergm: Hierarchical exponential-family random graph models', *Journal of Statistical Software* .

Sheng, Shuyang (2012), Identification and estimation of network formation games.

Snijders, Tom A.B (2002), 'Markov chain monte carlo estimation of exponential random graph models', *Journal of Social Structure* **3**(2).

Stephens, Matthew (2000), 'Dealing with label switching in mixture models', *Journal of the Royal Statistical Society B* .

Topa, Giorgio (2001), 'Social interactions, local spillovers and unemployment', *Review of Economic Studies* **68**(2), 261–295.

Appendix A. Additional theoretical results

A.1. **Perfect segregation meeting technology.** Let's consider the special case in which $\rho_b(g_{-ij}, z_i, z_j) = 0$ for any pair (i, j), that is the case of perfect segregation for the meething process. It is easy to show that when this is the case I can factorize the likelihood of observing the network as the product of K subnetwork likelihoods, one for each community.

PROPOSITION 2. ***Perfect segregation meeting technology.***
If $\rho_b(g_{=ij}, x_i, x_j) = 0$, then the likelihood of observing network g, conditional on the community structure z and covariates x is

$$\pi(g, x, z; \theta) = \prod_{k=1}^{K} \frac{\exp\left[Q_{k,k}(g_{k,k}, x^{(k)}, z; \theta)\right]}{c_{k,k}(\mathcal{G}_{k,k}, x^{(k)}; \theta)} \tag{39}$$

where the potential $Q_{k,k}(g_{k,k}, x^{(k)}, z; \theta)$ is

$$Q_{k,k}(g_{k,k}, x^{(k)}, z; \theta) = \sum_{i \in \mathcal{C}_k} \sum_{j \in \mathcal{C}_k} g_{ij} u(x_i, x_j, z_i, z_j; \alpha, \beta) + \frac{\gamma_k}{6} \sum_{i \in \mathcal{C}_k} \sum_{j \in \mathcal{C}_k} \sum_{r \in \mathcal{C}_k} g_{ij} g_{jr} g_{ri} \tag{40}$$

and the normalizing constant $c_{k,k}(\mathcal{G}_{k,k}, x^{(k)}; \theta)$ is

$$c_{k,k}(\mathcal{G}_{k,k}, x^{(k)}; \theta) = \sum_{\omega_{k,k} \in \mathcal{G}_{k,k}} \exp\left[Q_{k,k}(\omega_{k,k}, x^{(k)}, z; \theta)\right] \tag{41}$$

Proof. When $\rho_b(g_{-ij}, x_i, x_j) = 0$, there is no meeting of players across communities, therefore no links across community will be formed. Each community is independent; within the community, the network formation process is the same as Mele (2017) and Mele and Zhu (2017) (see Theorem 1 in Mele (2017) for a proof). Therefore each community subnetwork converges to an exponential random graph, with potential function $Q_{k,k}(g_{k,k}, x^{(k)}, z; \theta)$. □

The likelihood in (39) factorizes as a product of K independent and identically distributed exponential random graphs. The factorization is quite useful in estimation and for the identification of the parameters. In terms of estimation, I can parallelize the estimation routines, simulating a subnetwork on each processor. In addition, given the i.i.d. nature of the sample, I obtain identification with standard regularity conditions for the exponential family. In particular, as the number of communities grows large, the parameters are identified.[49]

Appendix B. Asymptotic Normality of sufficient statistics

This appendix contains the detailed proof of Theorem 3 in the main text. The model generates sparse networks with weak dependence among links. One can show that these features imply that the sufficient statistics are asymptotically normally distributed, as long as the number of communities is sufficiently large. To prove this result, I use Theorem 2 in Schweinberger and Handcock (2015). I will prove that all the conditions of their theorem are satisfied for my model. I report their result for completeness.

[49] See Lehman (1983) for a general discussion. See also Badev (2013).

THEOREM 4. ***(Theorem 2 in Schweinberger and Handcock (2015))*** *Let $\mathcal{C}_1, \mathcal{C}_2, ...$ be a sequence of non-empty, finite sets of nodes and $g_1, g_2, ...$ be a sequence of networks with increasing domain $N_1 \times N_1, N_2 \times N_2, ...$, with $N_K = \bigcup_{k=1}^{K} \mathcal{C}_k$. Let $\mathcal{S}_K \subseteq \times_{i=1}^{d} N_K$ be a subset of the d-dimensional Cartesian product of N_K with itself. Let S_K be a real-valued function with domain $\mathcal{S}_K$. Consider sums of the form $S_K = \sum_{s \in \mathcal{S}_K} S_{K,s}$, where $S_{K,s} = \prod_{k=1}^{q} g_{s,a_k,b_k}$ is the interaction of q distinct links g_{s,a_k,b_k}. Assume that the edge variables g_{ij} satisfy uniform boundedness in the sense that there exists a constant $C > 0$ such that, for all $K > 0$, $a \in N_K$ and $b \in N_K$, $\mathbb{P}(|g_{ab}| \leq C) = 1$. Without loss of generality, assume that, for all $K > 0$ and $i \in \mathcal{S}_K$, $\mathbb{E}(S_{K,s}) = 0$. If the sequence $g_1, g_2, ...$ is local and $\delta >$ d-sparse and $\mathbb{V}(W_K) \to \infty$ as $K \to \infty$, then*

$$\lim_{K \to \infty} \max_{1 \leq k \leq K} \mathbb{P}\left(|W_{K,k}| > \epsilon \sqrt{\mathbb{V}(W_K)} \right) = 0 \tag{42}$$

and

$$\frac{S_K}{\sqrt{\mathbb{V}(S_K)}} \xrightarrow{d} N(0,1) \text{ as } K \to \infty \tag{43}$$

In Theorem 4 the sums S_K are sufficient statistics of the exponential family distribution. For example, in our model when $q = 3$, the variables $S_{K,s} = g_{ab} g_{bc} g_{ac}$ are interactions of 3 links, therefore an indicator of whether c is a common friend of a and b. Thus $S_K = \sum_{s \in \mathcal{S}_K} S_{K,s} = \sum_i \sum_j \sum_r g_{ij} g_{jr} g_{ir}$. I can construct the other sufficient statistics in analogous fashion, using the notation in Theorem 4.

The first condition that needs to be satisfied is $\delta > 3$ in our model. This affects the speed at which the meetings of players of different communities can meet, as the size of the network n grows large. For the rest of the proof I will assume that this condition is satisfied.

I will prove that my model satisfies all the conditions in Theorem 4, by a series of Lemmas. Lemma 1 shows that I can decompose the sufficient statistics of our model in two components: within- and between-neighborhoods. Lemma 2 shows that the binary link variables satisfy the uniform boundedness condition. Lemma 3 proves that our model satisfies the sparsity requirement in Schweinberger and Handcock (2015). Lemma 4 proves that the network is local, that is the likelihood can be factorized in within- and between-communities components.

LEMMA 1. *The sufficient statistics of the model can be decomposed in within- and between-neighborhoods statistics.*

Proof. The first sufficient statistics is

$$S_{1K} := \sum_{i=1}^{n} \sum_{j=1}^{n} g_{ij} \tag{44}$$

We can decompose S_{1K} in two parts as follows

$$S_{1K} = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{1}_{\{z_i = z_j\}} g_{ij} + \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{1}_{\{z_i \neq z_j\}} g_{ij} = W_{1K} + B_{1K} \tag{45}$$

We can further decompose W_{1K} as

$$W_{1K} = \sum_{k=1}^{K} W_{1K,k} = \sum_{k=1}^{K} \sum_{i=1}^{n} \sum_{j=1}^{n} z_{ik} z_{jk} g_{ij} \tag{46}$$

The decomposition for the terms with covariance is trivial, while for triangles we obtain the following

$$S_{3K} := \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{r=1}^{n} g_{ij} g_{jr} g_{ri} \tag{47}$$

That decomposes as

$$S_{3K} = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{r=1}^{n} \mathbf{1}_{\{z_i = z_j = z_r\}} g_{ij} g_{jr} g_{ri} + \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{r=1}^{n} \mathbf{1}_{\{z_i \neq z_j \text{ or} z_i \neq z_r\}} g_{ij} g_{jr} g_{ri} = W_{3K} + B_{3K} \tag{48}$$

Notice that because of our assumptions on the meeting process, the second term B_{3K} will converge to zero asymptotically. □

LEMMA 2. *The variables g_{ij}'s satisfy uniform boundedness.*

Proof. We need to show that the link probabilities are bounded uniformly. The network is binary, so by choosing any $C > 1$, we have that for any K

$$\mathbb{P}\left(|g_{ij}| < C\right) = \mathbb{P}\left(g_{ij} < C\right) = 1 \tag{49}$$

for any i and j in the set of players. □

LEMMA 3. *Let $\mathcal{C}_1, \mathcal{C}_2, ...$ be a sequence of non-empty, finite sets of players and let $g_1, g_2, ...$ be a sequence of graphs from our model, with increasing domain $N_1 \times N_1, N_2 \times N_2, ...$, with $N_K = \bigcup_{k=1}^{K} \mathcal{C}_k$. Then there exist constants $A > 0$ and $\delta > 0$ such that for every pair i and j with $z_i \neq z_j$, we have $\mathbb{E}(|g_{ij}|^b) \leq A n^{-\delta}, b = 1, 2$.*

Proof. Since the network is binary we have $\mathbb{E}(|g_{ij}|^b) = \mathbb{E}(g_{ij})$ for $b = 1, 2$. Therefore we have

$$\mathbb{E}(g_{ij}) = Pr(\text{drawing } i, j | z_i \neq z_j) \times Pr(g_{ij} = 1 | \text{drawing } i, j) \tag{50}$$

$$= \frac{\rho(g_{-ij}, x_i, x_j)}{n^{\delta}} \frac{\exp\left[2u(\alpha, \beta, x_i, x_j, z_i, z_j)\right]}{1 + \exp\left[2u(\alpha, \beta, x_i, x_j, z_i, z_j)\right]} \tag{51}$$

$$\leq \frac{\rho_b(g_{-ij}, x_i, x_j)}{n^{\delta}} \leq \frac{\max_{i \in N_K, j \in N_K} \rho_b(g_{-ij}, x_i, x_j)}{n^{\delta}} = \frac{A}{n^{\delta}} \tag{52}$$

□

LEMMA 4. *The sequence of random graphs $g_1, g_2, ...$ is local, i.e. the likelihood factorizes in within- and between-communities components.*

$$P_\theta\left(G=g|Z=z,X=x\right) = \prod_{k=1}^{K} P\left(G_{k,k}=g_{k,k}|Z=z,X=x;\theta\right) \times \left[\prod_{l>k}^{K} P\left(G_{k,l}=g_{k,l}|Z=z,X=x;\theta\right)\right] \tag{53}$$

Proof. Let $g_{k,l}$ denote the subnetwork formed by individuals of communities $\mathcal{C}_k$ and $\mathcal{C}_l$. Let $x^{(k)}$ denote the covariates of players in community $\mathcal{C}_k$. The potential can be decomposed into the sum of sub-potentials for the sub-networks $g_{k,l}$'s. That is, we can decompose the potential $Q(g,x,z,\theta)$ as sum of subpotentials $Q_{k,l}(g_{k,l},x^{(k)},x^{(l)},z)$, separating the within-community and between-community contributions as follows:

$$Q(g,x,z;\theta) = \sum_{k=1}^{K} Q_{k,k}(g_{k,l},x^{(k);\theta},z) + \sum_{k=1}^{K}\sum_{l>k}^{K} Q_{k,l}(g_{k,l},x^{(k)},x^{(l)},z;\theta) \tag{54}$$

$$= \sum_{k=1}^{K}\left[\sum_{i\in\mathcal{C}_k}\sum_{j\in\mathcal{C}_k} g_{ij}u(x_i,x_j,z_i,z_j;\alpha,\beta) + \frac{\gamma_k}{6}\sum_{i\in\mathcal{C}_k}\sum_{j\in\mathcal{C}_k}\sum_{r\in\mathcal{C}_k} g_{ij}g_{jr}g_{ri}\right] \tag{55}$$

$$+ \sum_{k=1}^{K}\sum_{l>k}^{K}\left[\sum_{i\in\mathcal{C}_k}\sum_{j\in\mathcal{C}_l} g_{ij}u(x_i,x_j,z_i,z_j;\alpha,\beta)\right] \tag{56}$$

This decomposition of the potential function allows us to rewrite the likelihood as follows

$$\pi(g,x,z;\theta) = \prod_{k=1}^{K}\frac{\exp\left[Q_{k,k}(g_{k,k},x^{(k)},z;\theta)\right]}{c_{k,k}(\mathcal{G}_{k,k},x^{(k)};\theta)}\left[\prod_{l>k}^{K}\frac{\exp\left[Q_{k,l}(g_{k,l},x^{(k)},x^{(l)},z;\theta)\right]}{c_{k,l}(\mathcal{G}_{k,l},x^{(k)},x^{(l)};\theta)}\right] \tag{57}$$

$$= \prod_{k=1}^{K} P\left(g_{k,k}|z,x;\theta\right)\left[\prod_{l>k}^{K} P\left(g_{k,l}|z,x;\theta\right)\right] \tag{58}$$

where the normalizing constants $c_{k,k}(\mathcal{G}_{k,k},x^{(k)};\theta)$ and $c_{k,l}(\mathcal{G}_{k,l},x^{(k),x^{(l)}};\theta)$ are

$$c_{k,k}(\mathcal{G}_{k,k},x^{(k)};\theta) = \sum_{\omega_{k,k}\in\mathcal{G}_{k,k}}\exp\left[Q_{k,k}(\omega_{k,k},x^{(k)},z;\theta)\right] \tag{59}$$

$$c_{k,l}(\mathcal{G}_{k,l},x^{(k)},x^{(l)};\theta) = \sum_{\omega_{k,l}\in\mathcal{G}_{k,l}}\exp\left[Q_{k,l}(g_{k,l},x^{(k)},x^{(l)},z;\theta)\right] \tag{60}$$

Notice that the potential decomposition above, is consistent with the result in Lemma 1; that is, the sufficient statistics of the model can be written as sum of within- and between-communities sufficient statistics. □

Appendix C. Estimation details

All the computations have been performed on a desktop Dell Precision T7620 with 2 Intel Xeon CPUs E5-2697 v2 with 12 Dual core processors at 2.7GHZ each and 64GB of RAM.

I estimated all the models using the package `hergm` in R, developed by Schweinberger and Handcock (2015) and Schweinberger and Luna (forthcoming). The code for estimation and replication is available from the author.

Each estimate is obtained with a 100,000 simulation run of the exchange algorithm. I collect 10,000 samples and discard the first 2,000 as burnin. I also experimented with a longer run of 200,000 steps, without changes in the results.

Add Health restricted-use data used in this paper can be obtained by applying at the website: http://www.cpc.unc.edu/projects/addhealth

C.1. **Prior truncation.** In the empirical application, the priors are truncated, so that the number of communities is at most K_{max}. Therefore we have

$$\eta_1 = V_1 \tag{61}$$

$$\eta_k = V_k \prod_{j=1}^{K_{max}-1} (1 - V_j) \qquad k = 2, 3, 4, ..., K_{max} \tag{62}$$

$$V_k|\phi \overset{iid}{\sim} Beta(1, \phi) \qquad k = 1, 2, 3, ..., K_{max} - 1 \tag{63}$$

$$V_{K_{max}} = 1 \tag{64}$$

$$\phi > 0 \quad \text{and} \quad \sum_{k=1}^{K_{max}} \eta_k = 1 \ w.p.1 \tag{65}$$

This simpler formulation with truncation provides a more parsimonious model and improves the speed of computations and simulations from the posterior. Truncation is indeed needed because the number of parameters to estimate depends on the number of communities and we have only one network observation.

C.2. **Hyper-priors used in estimation.** I use the default in the `hergm` package in R. The hyper prior on ϕ is

$$\phi \sim Gamma(1, 1) \tag{66}$$

(67)

The hyper-priors for $\mu_w, \mu_b, \sigma_w, \sigma_b$ are

$$\mu_w \sim N(0, 1) \tag{68}$$

$$\mu_b \sim N(0, 1) \tag{69}$$

$$\sigma_w \sim Gamma(10, 10) \tag{70}$$

$$\sigma_b \sim Gamma(10, 10) \tag{71}$$

(72)

All the variables are independent.

C.3. **Posterior estimates for $K_{max} = 5$.** In Table 2 we report the estimated posterior for the structural parameters, when the prior is truncated at a maximum of 5 communities ($K_{max} = 5$). The estimates for the costs of links are less precise than in the case of $K_{max} = 3$.

The same is true for the payoffs from common friends and the parameters for homophily in covariates. The results are qualitatively the same as the ones in Table 1.

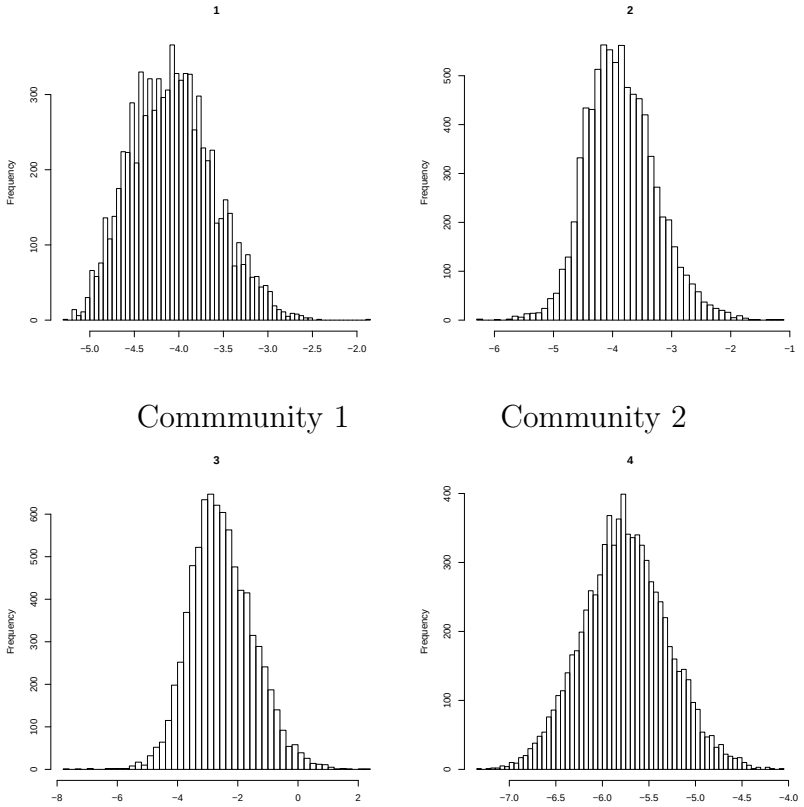
TABLE 2. Estimates of the posterior for $K_{max} = 5$.

Parameter	Post. Mean	Post. s.d.	Posterior quantiles 2.5%	50%	97.5%
A. Cost of links					
α_1	-3.967	0.530	-4.898	-3.958	-2.951
α_2	-2.754	1.132	-4.810	-2.839	-0.252
α_3	-2.747	1.012	-4.784	-2.761	-0.606
α_4	-2.525	1.119	-4.702	-2.550	-0.209
α_5	-2.596	1.190	-4.757	-2.654	-0.105
α_b	-5.474	0.533	-6.485	-5.464	-4.521
B.Payoff from covariates					
$\beta_{white,white}$	0.889	0.313	0.302	0.867	1.447
$\beta_{black,black}$	0.834	0.297	0.213	0.831	1.394
$\beta_{hisp,hisp}$	1.852	0.760	-0.415	1.923	2.992
$\beta_{grade7,grade7}$	1.387	0.335	0.618	1.397	1.962
$\beta_{grade8,grade8}$	1.260	0.406	0.433	1.290	1.951
$\beta_{grade9,grade9}$	1.159	0.405	0.350	1.182	1.900
$\beta_{grade10,grade10}$	1.009	0.498	-0.247	1.035	1.838
$\beta_{grade11,grade11}$	1.270	0.497	0.181	1.307	2.082
$\beta_{grade12,grade12}$	0.993	0.417	0.107	1.016	1.742
$\beta_{male,male}$	-0.180	0.375	-0.984	-0.143	0.530
$\beta_{female,female}$	-0.319	0.297	-0.880	-0.326	0.268
$\beta_{\lvert income_i - income_j \rvert}$	-0.695	0.291	-1.280	-0.685	-0.145
C. Payoff from common friends					
γ_1	1.007	0.168	0.635	1.016	1.311
γ_2	1.058	1.087	-1.270	1.102	3.120
γ_3	0.724	1.103	-1.516	0.753	2.822
γ_4	0.709	1.092	-1.522	0.691	2.961
γ_5	0.759	1.148	-1.522	0.777	2.926

C.4. **Posterior estimates.** Histograms of the marginal posteriors for the cost of links α_k's are shown in Figure 4. The marginal posteriors for the covariates are shown in Figure 5. The marginal posterior of the preference for common friends γ_k's are in Figure 6.

C.5. **Posterior predictions on community structure.** In Figure 7, we show the posterior estimates of community structure for the different values of K_{max}. The esitimated membership do not differ too much for $k = 3$ and higher. So this confirms our choices of $K_{max} = 3$ as main specification of interest.

FIGURE 4. Estimated marginal posterior of cost of links for different communities (α)



Commmunity 1 Community 2

Commmunity 3 Between communities

JOHNS HOPKINS UNIVERSITY, CAREY BUSINESS SCHOOL, 100 INTERNATIONAL DR, BALTIMORE, MD 21202

FIGURE 5. Estimated marginal posterior of preference parameter for observable characteristics (β)

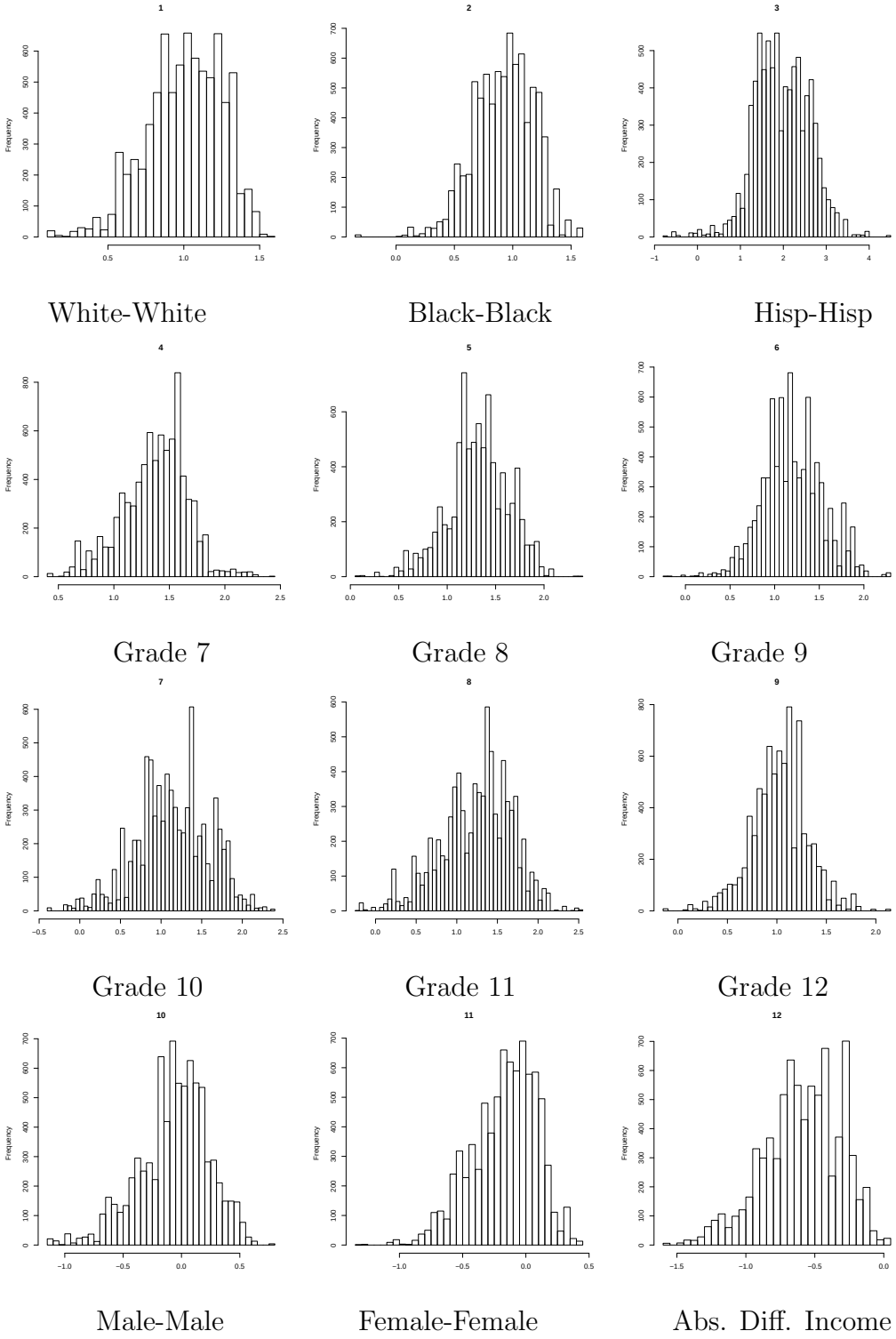
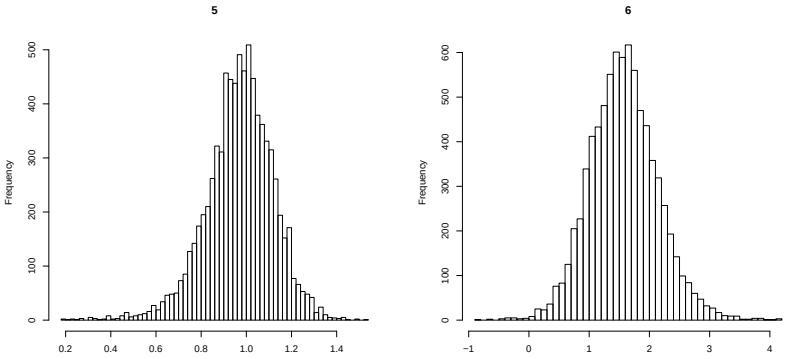
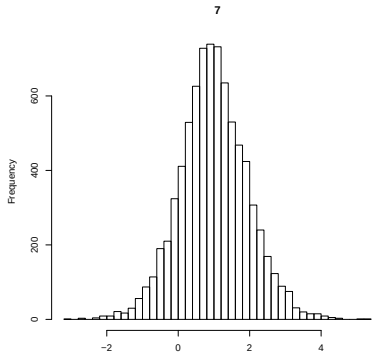


FIGURE 6. Estimated marginal posterior of preference parameter for common friends in different communities (γ)

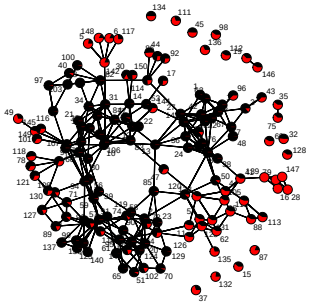


Commmunity 1 Community 2

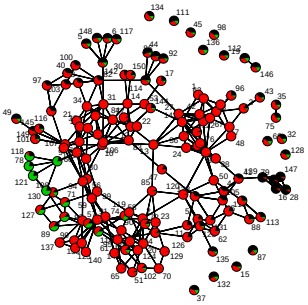


Commmunity 3

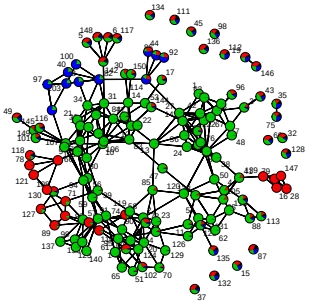
FIGURE 7. Posterior predictions for community structure



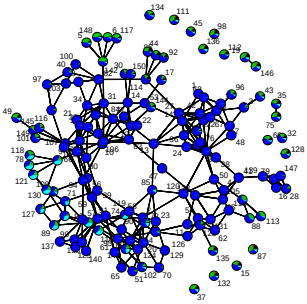
2 communities



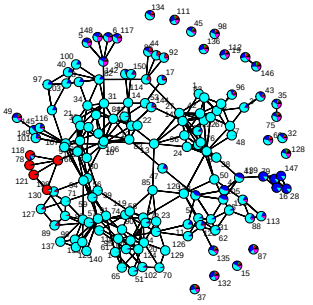
3 communities



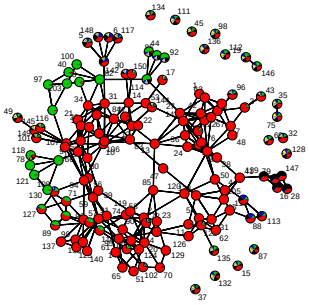
4 communities



5 communities



6 communities



10 communities