Inferring social kernels and segregation measures from ego-network data

Till Hoffmann and Nick S. Jones

Department of Mathematics, Imperial College London

How people connect with one another is a fundamental question in the social sciences, and the resulting social networks can have a profound impact on our daily lives. Blau offered a powerful explanation: people connect with one another based on their positions in a social space. Yet a principled measure of social distance in society remains elusive.

We use the connectivity kernel of conditionally-independent edge models to study distance in Blau space and develop a family of segregation measures with desirable properties: they have an intuitive and universal scale, are applicable to multivariate and mixed node attributes, and capture segregation at the level of individuals, pairs of individuals, and society as a whole. We use the measures to provide visual maps of Blau space—a space spanned by the attributes of the members of society.

Under a Bayesian paradigm, we infer the parameters of the connectivity kernel from eleven ego network datasets collected in four surveys in the United Kingdom and United States. The importance of different dimensions of Blau space is similar across time and location, suggesting a macroscopically stable social fabric. Physical separation and age differences have the most significant impact on segregation within friendship networks with implications for intergenerational mixing and isolation in later stages of life.

1. Introduction

Homophily, the tendency for people to connect with others who are alike, is one of the most robust observations of the social sciences and shapes how our society is connected [1]. Quantifying homophily is not only important for understanding why social ties form between some people yet not between others, but the manifestation of homophily as poorly-connected social networks can have a significant impact on dynamics unfolding upon them [2]. For example, users of online social networks, such as Facebook and Twitter, tend to connect with others who hold similar political views [3]. They are more likely

to be exposed to information that confirms rather than challenges their beliefs [4]. An "echo chamber" effect ensues, leading to polarised opinions [5]. Homophily can also have a detrimental impact on public health: clusters of individuals who mutually reinforce their belief that vaccinations are harmful can raise the likelihood of significant disease outbreaks [6]—even if the vaccination rate is above herd immunity levels on average.

Homophily can be observed in friendships [7, 8], networks of discussion partners [9], communication networks [10, 11], marital ties [12], and online social networks [13]. Relationships are homogeneous with respect to a wide range of attributes, including age [14, 15], gender [15, 16], ethnicity [7, 13, 17], education [9, 15, 18], occupation [19], income [10, 11, 18], religion [20], parental [18] and marital status [21], political ideology [3, 4], and geographical location [22–26].

Segregation measures are often used to quantify homophily [27, 28]. Many approaches are based on co-presence in organisational units such as schools [29], voluntary associations [30], occupations [31], or census tracts [32, 33], and we refer to them as *organisational* measures. Typically, they compare how the distribution of demographic attributes within organisational units differs from the distribution of attributes in the general population. Whilst organisational measures are applicable whenever data can be stratified according to a variable of interest, they cannot capture segregation at smaller scales than the strata [17]. For example, the ethnic composition of a set of schools may be representative of the general population, indicating that there is no (organisational) segregation. But the social networks *within* schools often exhibit strong ethnic homophily [7, 34]. Organisational measures cannot capture such social segregation.

Social measures of segregation, such as the assortativity coefficient [35], overcome these limitations by explicitly considering the interactions amongst individuals [17], but have their own difficulties: first, they usually rely on the existence of mutually exclusive groups such as sex, ethnicity, or religion [28], and they are not applicable to continuous attributes such as age or income. Attributes are often discretised [21, 36, 37], but the boundaries between categories always suffer from some degree of arbitrariness [32]. Second, segregation for multiple attributes can be measured independently, but it is unclear how to define a composite segregation measure. Furthermore, social measures are typically defined as summary statistics of a fully-observed social network. Consequently, we cannot easily quantify uncertainties.

In practice, the study of homophily is complicated by the scarcity of high-quality data [17, 38]: we need social network data together with demographic information for each person. Online social networks and the widespread use of mobile phones provide us with detailed information about connections between individuals [39], and seemingly private traits such as socioeconomic status [40, 41], sexual orientation [42], age, gender, and political ideology can be inferred [43]. Unfortunately, network features are often used to predict demographic attributes [10, 40, 41, 43], which would confound any study of homophily. Furthermore, data are too revealing in terms of privacy but, at the same time, do not provide enough information for researchers [39]. Individuals can be identified in anonymised social networks [44, 45], and augmenting the network data with demographic information would make re-identification even easier.

However, censuses and large-scale surveys collect comprehensive demographic inform-

ation from respondents but usually lack data about their associates. Fortunately, some surveys have included questions about respondents' friends [18, 46], discussion partners [9, 47], or support networks [21, 48]. The questions used to elicit social ties provide an imperfect observation of the immediate neighbourhood of respondents [49–51].

Building on the successes of conditionally-independent edge models [52] and, in particular, latent space models for social networks [53–55], we consider a generative model for social networks whose members occupy a multidimensional Blau space. We discuss desirable properties for social segregation measures, and, using the generative network model, we develop a suite of measures applicable to arbitrary attributes. The measures capture segregation at different scales: single individuals, pairs of individuals, and society as a whole. We illustrate the measures with a simple example, and we show that it reduces to a well-known segregation measure if the attributes are univariate and categorical.

Furthermore, we derive the posterior for parameters of the conditionally-independent edge model given partial observations of social networks obtained from surveys. We apply our approach to nine existing datasets from the United Kingdom and two from the United States. Our analysis reveals that the effects of homophily on society are remarkably stable in both countries regardless of time and the specific nature of relationships. Using the suite of segregation measures, we find that physical separation and age are the most important factors contributing to the segregation of society. We provide recommendations for conducting surveys to infer homophily in social networks and discuss future work, including principled imputation of demographic attributes in social networks and using the generative model to generate synthetic networks: an approach that has attracted interest to circumvent the difficulties associated with anonymising social network data [56–60].

2. Methods

2.1. Generative network model

We consider a generative model for social networks for a population of n individuals N who occupy a Blau space $\mathbb{B}$ spanned by their demographic attributes, such as age, income, or gender. In contrast to common latent space models [53, 54, 61], the attributes are observed, although Hoff, Raftery and Handcock [53] also consider an extension including covariates. The q-dimensional attribute vector $x_i \in \mathbb{B}$ for each individual $i \in N$ is drawn independently from a distribution $P(x_i)$ of demographic attributes. Elements of the attribute vector can take continuous, ordinal, or categorical values. Connections between individuals are encoded by the binary adjacency matrix A such that $A_{ij} = 1$ if j considers i to be a friend and $A_{ij} = 0$ otherwise. We assume that people do not interact with themselves such that $A_{ii} = 0$ for all i, and that connections are undirected, although social ties need not be reciprocated in general [62].

Given the positions of two individuals i and j in Blau space, we assume that connections form independently with probability $\rho(x_i, x_j)$, i.e. edges are conditionally-independent given the attributes of nodes [63]. The assumption of conditionally-indepen-

dent edges can be problematic. For example, it is not possible to reproduce heavy-tailed degree distributions if the node density is homogeneous [64]. Furthermore, the average degree scales linearly with the number of nodes unless the connectivity kernel ρ is adjusted to compensate [65]. Nevertheless, we use spatial network models because the connectivity kernel is intuitive, and they can capture salient features of social networks. For example, nodes in high-density regions have larger degrees on average [64]. Similarly, members of the ethnic majority have more social ties in social networks in US high schools [7].

2.2. Developing a model-based segregation measure

In addition to addressing the challenges mentioned in section 1, a social segregation measure should satisfy the following properties: first, the measure should be insensitive to the overall edge density to facilitate comparison of segregation across different networks. Otherwise, the segregation measure would depend on the size of the population because the edge density scales as n^{-1} if the average degree is approximately constant. Second, following Freeman [66], we would like the measure to capture the notion that segregation places "restrictions on the access of people to one another". Third, the measure should be easily interpretable, and it should have a natural notion of the absence of segregation when individuals form connections without regard to their positions in Blau space. For example, the difference of within- and between-group ties considered by Krackhardt and Stern [67] depends on the sizes of the groups even if there is no homophily: there is no natural reference point.

A single measure cannot capture the complexities of social networks, and we develop a family of measures applicable at different scales: the social separation between any two individuals, the isolation experienced by any one individual, and the social strain experienced by society as a whole. Starting at the microscopic level, we use the generative model to define a measure of *social separation* between two individuals with attributes x and y:

$$\varphi(x, y) = \operatorname{logit} \rho(y, y) - \operatorname{logit} \rho(x, y), \tag{1}$$
$$\text{where } \operatorname{logit} \rho = \log\left(\frac{\rho}{1-\rho}\right)$$

are the log odds for a connection to form with probability ρ. The probability $\rho(y, y)$ for an individual with attributes y to connect with someone with identical attributes serves as a reference point, and the measure does not depend on the overall edge density. The social separation φ may be understood as the isolation experienced by an individual with attributes x as a result of the behaviour of another with attributes y. The measure is zero if people do not discriminate with respect to the attributes on which they differ or if two individuals have the same demographic attributes. For a homophilous connectivity kernel, the measure is positive and is a *semi-metric* for Blau space; we will consider a family of connectivity kernels for which φ is a true metric in section 2.3.

Proposition 1. *If the connectivity kernel is homophilious, symmetric, and the probability $\rho(x, x)$ to connect with others who are alike is independent of x, the social separation φ is a* semimetric *[68]: it satisfies the properties of a metric, including non-negativity, symmetry, and the identity of indiscernables—except the triangle inequality.*

Proof. First, $\varphi(x, y) \geq 0$ because homophily implies that $\rho(x, y) < \rho(y, y)$ and logit is a monotonically increasing function. Second, $\varphi(x, y) = \varphi(y, x)$ because the first term of eq. (1) is constant by assumption and the second is symmetric because the kernel is symmetric. Third, the measure is zero for any two individuals with the same attributes by substitution into eq. (1). Similarly, if $\varphi(x, y) = 0$, then $x = y$ because $\rho(x, y) < \rho(y, y)$. □

The less likely two people are to connect, the larger the social separation between them. The assumptions required for proposition 1 to hold may seem restrictive, but they are satisfied by most studies of spatial networks [22, 23, 38, 64].

Defining social separation in terms of a generative model, i.e. using the connectivity kernel rather than a summary statistic of a particular dataset, provides us with two advantages: first, any uncertainty associated with the inferred connectivity kernel naturally propagates to the segregation measure. Second, we can easily consider the properties of the segregation measure under a variety of generative models without having to resort to computationally-expensive Monte Carlo simulations.

For example, consider a stochastic block model (SBM) [52] with K blocks, intra-group connection probability ρ_{same}, and inter-group connection probability $\rho_{\text{different}} < \rho_{\text{same}}$. For two nodes with block membership x and y, the social separation is

$$\varphi(x, y) = (1 - \delta_{xy}) \left(\operatorname{logit} \rho_{\text{same}} - \operatorname{logit} \rho_{\text{different}} \right), \tag{2}$$

δ_{xy} is the Kronecker delta. The social separation only depends on block membership, and it is not affected by the size of each block. For members of different blocks, φ is the difference of log odds ratios for the existence of intra-group ties as opposed to inter-group ties previously considered by Moody [34].

To quantify segregation at the level of an individual with attributes x, we define the *social isolation*

$$\phi(x) = \int dy \; P(y) \varphi(x, y), \tag{3}$$

which measures the average social separation between an individual with attribute x and members of society whose attribute distribution is $P(y)$. For the SBM, we substitute eq. (2) into eq. (3) and obtain

$$\phi(x) = (1 - P(x)) \left(\operatorname{logit} \rho_{\text{same}} - \operatorname{logit} \rho_{\text{different}} \right), \tag{4}$$

where $P(x)$ is the probability to belong to block x. Members of all blocks experience the same degree of isolation if the blocks are of the same size. If the sizes are unequal, minorities experience more isolation and majority groups experience less isolation. Indeed, ethnic minorities in schools tend to be more isolated and have fewer social ties [7].

To understand how segregated society is as a whole, we would like to aggregate the social isoluation ϕ, but the appropriate statistic depends on the question at hand. For example, if we wanted to study the most isolated subpopulation of society, we should consider $\max_{x\in\mathbb{B}} \phi(x)$. Here, we take a utilitarian approach and, in line with eq. (3), define the *social strain* as

$$\Phi = \int dx\ P(x)\phi(x), \tag{5}$$

which quantifies the average social separation amongst members of the society. It is zero when individuals do not discriminate based on attributes, and it can reach arbitrarily large values in a society comprising multiple groups that are completely disconnected. For the SBM, we substitute eq. (4) into eq. (5) and obtain

$$\begin{aligned} \Phi &= \gamma \left(\operatorname{logit} \rho_{\text{same}} - \operatorname{logit} \rho_{\text{different}}\right), \\ \text{where } \gamma &= \sum_{x=1}^{K} P(x)\left(1 - P(x)\right) \end{aligned} \tag{6}$$

is a factor accounting for the relative sizes of the K blocks. The social strain is maximal when the group sizes are equal. If one of the blocks is larger, the strain approaches zero as the sizes of the minority blocks decrease: members of the majority group experiences little social isolation. It is unsurprising that there is no social strain if the society is homogeneous, but the utilitarian approach has a serious limitation: it has little concern for minorities that are not well integrated in society. For equal group sizes, the social strain increases with the number of groups, asymptotically reaching a maximum value of $\operatorname{logit} \rho_{\text{same}} - \operatorname{logit} \rho_{\text{different}}$.

2.3. Distance in Blau space

The social separation takes a simple form if the probability for two individuals to connect is a logistic kernel [53], i.e.

$$\operatorname{logit} \rho(x, y, \theta) = \sum_{l=1}^{p} \theta_l f_l(x, y), \tag{7}$$

where the p-dimensional vector θ_l parametrises the kernel, and $f(x, y)$ is a set of p-dimensional features that are predictive of the connection probability, such as the age difference $f_{\text{age}} = |x_{\text{age}} - y_{\text{age}}|$. The social separation between x and y comprises contributions from the features of the logistic kernel:

$$\varphi(x, y) = \sum_{l=1}^{p} \varphi_l(x, y), \tag{8}$$

$$\text{where } \varphi_l(x, y) = \theta_l \left(f_l(y, y) - f_l(x, y)\right) \tag{9}$$

is the contribution due to a single feature l. In fact, φ is a true metric for many logistic connectivity kernels.

Proposition 2. *The social separation $\varphi(x, y)$ is a metric if the kernel is homophilous, i.e. $\theta_l < 0$, and each feature $f_l(x, y)$ is a constant or a positive affine transform of a metric $d_l(x, y)$, i.e.*

$$f_l(x, y) = a_l d_l(x, y) + b_l, \tag{10}$$

where $a_l > 0$ and b_l are the parameters of the affine transform.

Proof. The social separation $\varphi(x, y)$ is a semi-metric according to proposition 1, and, according to eq. (8), it comprises contributions from individual features. Showing that each contribution $\varphi_l(x, y)$ satisfies the triangle inequality is sufficient for $\varphi(x, y)$ to satisfy it, i.e. we require

$$\varphi_l(x, z) \geq \varphi_l(x, y) + \varphi_l(y, z). \tag{11}$$

Substituting eq. (10) into eq. (11) yields

$$-\theta_l a_l d_l(x, z) \geq -\theta_l a_l \left[d_l(x, y) + d_l(y, z)\right], \tag{12}$$

where we have used the metric property $d_l(x, x) = 0$ for all x, and the constant b_l in eq. (10) vanishes by eq. (9). The inequality in eq. (12) holds because $\theta_l < 0$ for homophilous kernels, $a_l > 0$ by assumption, and $d_l(x, y)$ is a metric. Equation (11) is trivially satisfied for a constant feature, such as a bias term controlling the overall edge density. □

In other words, the social separation is a true measure of *distance* in the social space with a probabilistic interpretation if features are themselves measures of distance, including all the features we consider subsequently. This observation puts Peter Blau's [69] hypothesis that "the macrostructure of societies can be defined as a multidimensional space of social positions among which people are distributed and which affect their social relations" on a sound statistical footing: *fitting conditionally-independent edge models allows us to learn the metric of Blau space.* The metric has a universal scale, facilitating comparison across different datasets.

2.4. Parameter inference given ego network data

A representative sample of dyads between individuals together with their demographic attributes is not generally available. However, a number of surveys have collected information about the social ties of respondents using name-generator questions which elicit social ties by asking respondents to nominate their friends [21], individuals they feel close to [8], or discussion partners [9, 47]. To generate examples of disconnected dyads, we consider a random sample of pairs of individuals. To account for this non-ignorable data collection process, we introduce a variable $I_{ij} \in \{0, 1\}$ indicating whether a particular dyad A_{ij} was observed. The available data thus comprise the demographic attributes x of individuals included in the sample and the dyad state A_{ij} (1 if i and j are connected and 0 otherwise) if it was observed, i.e. $I_{ij} = 1$. Adapting the argument presented by King and Zeng [70] to a Bayesian paradigm, we consider the posterior distribution over kernel parameters θ given the available data:

$$P(\theta|A, f, I = 1) \propto P(A|\theta, f, I = 1)P(\theta), \tag{13}$$

where $P(\theta)$ is the kernel parameter prior, and $f = f(x, y)$ are features sufficient to evaluate the connectivity kernel given demographic attributes x and y. The observed-data likelihood is

$$P(A|\theta, f, I = 1) = \frac{P(f|A, \theta, I = 1)P(A|\theta, I = 1)}{P(f|\theta, I = 1)}. \tag{14}$$

Considering the first term in the numerator of eq. (14), we note that the distribution over kernel features given the state A of the dyad does not depend on whether it was included in the sample or not. More formally,

$$P(f|A, \theta, I = 1) = P(f|A, \theta) \tag{15}$$

$$= \frac{P(A|f, \theta)P(f|\theta)}{P(A|\theta)}. \tag{16}$$

Turning to the denominator in eq. (14), we find

$$\begin{aligned} P(f|\theta, I = 1) &= \sum_{\alpha=0}^{1} P(f|A = \alpha, \theta, I = 1)P(A = \alpha|\theta, I = 1) \\ &= P(f|\theta) \sum_{\alpha=0}^{1} P(A = \alpha|f, \theta) \frac{P(A = \alpha|\theta, I = 1)}{P(A = \alpha|\theta)}, \end{aligned} \tag{17}$$

where we used the identity in eq. (16) to arrive at the second line. Substituting eqs. (16) and (17) into eq. (14), the observed-data likelihood is

$$P(A|\theta, f, I = 1) = \frac{P(A|f, \theta)r(A)}{\sum_{\alpha=0}^{1} P(A = \alpha|f, \theta)r(\alpha)}, \tag{18}$$

$$\text{where } r(\alpha) = \frac{P(A = \alpha|\theta, I = 1)}{P(A = \alpha|\theta)}$$

is the ratio of prevalences of dyad state α in the sample and the general population. In practice, we approximate the prevalence ratio r using the empirical sample prevalence and prior knowledge about the prevalence in the population. The posterior can be evaluated by substituting eq. (18) into eq. (13), and we can thus infer the parameters θ from ego network data. See appendix A.2 for details on how to evaluate the observed-data log-likelihood in a numerically stable fashion.

3. Application

3.1. Ego network data collected in surveys

The social ties identified through name-generator questions depend on the nature of the relationship, the mode of administration of the questionnaire (e.g. face-to-face, telephone interview, or online survey), and the interviewer [49, 50]. Consequently, we do not expect the kernel parameters inferred from different datasets to be completely consistent. In the

following investigation of ego networks, we restrict the nature of relationships to friends who are not relatives as much as the available data permit: we are interested in *voluntary* association amongst members of the population rather than the social structures they were born into [21].

Demographic information about nominees can be collected either by asking seeds about their friends' demographic background [9, 47] or by conducting follow-up surveys [18]. The latter seems preferable because respondents may not have complete information about their social contacts, but the approach requires additional resources to interview the nominees and may suffer from low response rates. For example, the age of nominees in the British Household Panel Survey (BHPS), one of the datasets we consider, is 60% more likely to be an integer multiple of ten than it is for seeds—presumably because seeds round the age of their friends to the nearest decade. In anticipation of such challenges, the coding for the nominees is often coarser than for seeds. To compare the demographic attributes of seeds and nominees we need to unify the coding (see appendix B for details for each dataset).

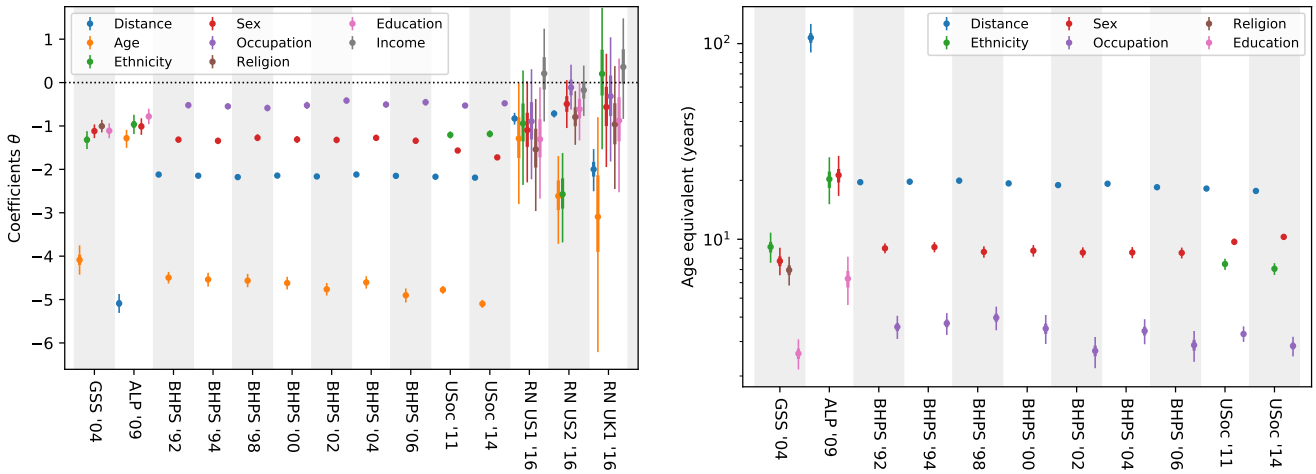
3.2. General Social Survey

The General Social Survey (GSS) is a nationally-representative face-to-face survey of non-institutionalised adults living in the US. Demographic attributes of seeds are collected regularly and include age, sex, ethnicity, religion, and education [14, 47]. In 2004, respondents were asked about the demographic background of people "with whom they discuss important matters", which tends to elicit close ties [49]. We omit all nominees who are not considered to be friends or who are family. The coding of age and sex is consistent amongst seeds and nominees. We aggregate the detailed coding of ethnic and religious attributes of seeds to match the coding of nominees, as shown in table B.1. Kernel features include the absolute age and ordinal education level difference as well as binary indicators for differences along the sex, ethnicity, and religion dimensions.

Some of the demographic attributes of seeds and nominees are missing because respondents did not know or refused to provide the information, and we drop dyads associated with individuals with one or more missing attributes, as shown in table B.2. Such a complete-case analysis can introduce biases if the data are not missing completely at random [71], but handling the missing data in a principled fashion would require us to develop a model for demographic attributes [72, 73].

For each demographic attribute, we define a feature for the logistic kernel in eq. (7), as shown in table B.1. To standardise the features $f(x_i, x_j)$, we subtract their mean and divide non-binary features by twice their standard deviation [74]; binary features are not rescaled. The statistics are calculated with respect to a random sample of pairs of seeds. Feature standardisation allows us to compare kernel parameters more easily [74] and simplifies the formulation of priors: we use independent, weakly-informative Cauchy priors for the kernel parameters such that

$$P(\theta_l) \propto \left[1 + \left(\frac{\theta_l}{\alpha_l}\right)^2\right]^{-1}.$$



[TAH: need to update figures]

Figure 1: *Age and physical separation have a strong impact on connection probabilities, and converting parameters into age equivalents makes feature comparison more intuitive.* Panel (a) shows kernel parameters inferred from ego network data for each dataset. Panel (b) shows age equivalents. For binary features (sex, occupation, religion, ethnicity, and distance for the American Life Panel), the equivalent number of years corresponds to a change from having the same attribute to having a different attribute. Age equivalents for the American Life Panel are overestimated (see section 3.3 for details). Markers represent the posterior mean, thick error bars the interquartile range, and thin error bars the 95% posterior interval.

Following Gelman et al. [75], we chose the scale parameters $\alpha_l = 2.5$ for $l > 1$ to represent our weak prior belief that changing a feature by one standard deviation is unlikely to change the log odds by more than five: the independent Cauchy distributions regularise the kernel parameters by placing significant prior probability near zero, but their heavy tails allow for significant departures from zero should the data be in support of large parameters. We set $\alpha_1 = 10$ because the parameter θ_1 associated with the constant bias term could change significantly depending on the population size [73, chapter 16].

The inference is performed in two steps: first, we maximise the posterior with respect to the parameters θ using a gradient ascent algorithm. Second, we run a Metropolis-Hastings algorithm to draw samples from the posterior [76]. Summary statistics of the posterior are shown in fig. 1 (a). The connection probabilities decrease quickly with increasing age differences. Ethnic, sex, educational, and religious differences all seem to have a similar effect and decrease the odds of connection by a factor of about 0.4 each.**[TAH: double check numbers]**

Hipp and Perrin [8] used the logarithm of physical separation as a benchmark to translate the effect of other attributes into distance-equivalents. We instead use age as

a benchmark because age is available for most datasets and is typically coded uniformly in years. In contrast, physical separation is often not available or coded heterogeneously across different datasets. For example, the American Life Panel only provides location information at the state level (see section 3.3), whilst the British Household Panel Survey recorded distance between seeds and nominees as ordinal data (see section 3.4). For the GSS, being of a different ethnicity is equivalent to a nine-year age difference, and having a different sex or religion translates to eight and seven years, respectively. One educational level, as defined in table B.1, corresponds to three years, as shown in fig. 1 (b).

3.3. American Life Panel

The American Life Panel (ALP) is a nationally-representative panel of adults resident in the US. Panel members are interviewed either using their own internet connection or are provided with a web television to access surveys. Data are collected regularly and each survey has a different focus. In 2009, information about social networks and financial literacy was collected. Demographic attributes included sex, age, ethnicity, education, their state of residence, and whether respondents identified as Hispanic. Respondents were also asked to nominate others with whom they "discuss financial matters" [77]. We only include nominees who are friends of seeds and exclude kinship ties; see table B.3 for details of harmonisation of attributes across seeds and nominees.

Homophily with respect to sex, ethnicity, and education is slightly weaker but not inconsistent with the GSS. Age differences appear to play less of a role in the discussion of financial matters at first sight, but the inference is severely biased for age. We cannot resolve strong age homophily because data are only recorded in 15-year bins: the small age parameter is likely a result of regression dilution caused by measuring ages imprecisely [78]. Consequently, the age equivalents in fig. 1 (b) are inflated. Being resident in a different state has by far the most significant impact on friendship formation.

3.4. British Household Panel Survey and Understanding Society

The British Household Panel Survey (BHPS) was a nationally-representative face-to-face survey in the UK. It was conducted from 1991 to 2008 and has since been replaced by the Understanding Society (USoc) survey. Respondents were asked questions about "their closest friends" every other year as part of the BHPS and every three years in USoc. Data include sex, age, occupational status, ethnicity (only in USoc), and how far away friends live [79, 80] (see table B.4 for details).

The inferred kernel parameters are largely consistent with the inference for the ALP and GSS in the US suggesting that friendship formation proceeds similarly in the two countries. We have omitted data from the BHPS in 2008 because we identified errors in the coding which have since been confirmed by the Institute of Social and Economic Research [81]. Similarly, we omitted data from the BHPS in 1996 because physical separation between friends was not recorded. As shown in fig. 1, homophily with respect to sex seems to have increased in recent years but it is unclear whether this observation is the result of people forming friendships differently or a different data collection

process [82].

3.5. Inferred segregation

To get a better understanding of Blau space and the metric induced by the connectivity kernel, we consider a sample S of 1,000 respondents from the General Social Survey and compute the social separation between them to obtain a distance matrix

$$\hat{\varphi}_{ij} = \hat{\theta}^{\mathsf{T}} \left(f(x_j, x_j) - f(x_i, x_j) \right),$$

where $\hat{\theta}$ is the posterior mean of the kernel parameters discussed in section 3.2. We use multidimensional scaling to embed the respondents in a two-dimensional space [83], as shown in fig. 2. Panel (b) shows the two-dimensional embedding that best approximates the distance matrix in the high-dimensional social space. Some men (squares) can be found amongst the cluster of women (circles) and vice versa because the optimisation algorithm converged to a local minimum.

The first dimension captures the age of respondents, as illustrated in panel (a): the mean age increases monotonically as a function of the first embedding dimension, and the standard deviation is small. We evaluated both statistics using Gaussian kernel smoothing [84, chapter 6]. The second dimension captures sex and ethnicity. As expected from eq. (4), ethnic minorities are more isolated and live on the outskirts of society while the ethnic majority occupies the centre. The embedding suggests that age has the strongest impact on how people form friendships.

Panel (b) of fig. 2 also shows the social isolation ϕ experienced by individuals as a greyscale heat map which we obtained in two steps: first, we evaluated an estimate of the social isolation

$$\hat{\phi}_i = \frac{1}{|S| - 1} \sum_{j \in S : j \neq i} \hat{\varphi}_{ij}.$$

Second, we applied Gaussian kernel smoothing to the social isolation in the embedding space. Respondents occupying the centre of society experience little isolation whereas individuals in the periphery are more isolated.

Similar to the social separation in eq. (8), the social strain can be broken down into components

$$\Phi_l = \theta_l \int dx \; dy \;\; (f_l(y, y) - f_l(x, y)) \tag{19}$$

because it is a linear functional of φ: each component contributes to the social strain in society. For each of the datasets, we evaluate an estimate of the contributions to the social strain

$$\hat{\Phi}_l = \frac{2\hat{\theta}_l}{|U| \left(|U| - 1 \right)} \sum_{i < j \in U : i \neq j} \left[f_l(x_j, x_j) - f_l(x_i, x_j) \right],$$

where the sum is over all distinct pairs of seeds U. The contribution Φ_l measures the average social separation due to feature l, and it captures the effect of both the connectivity kernel $\rho(x, y)$ and the attribute distribution $P(x)$: neither is sufficient on its own to

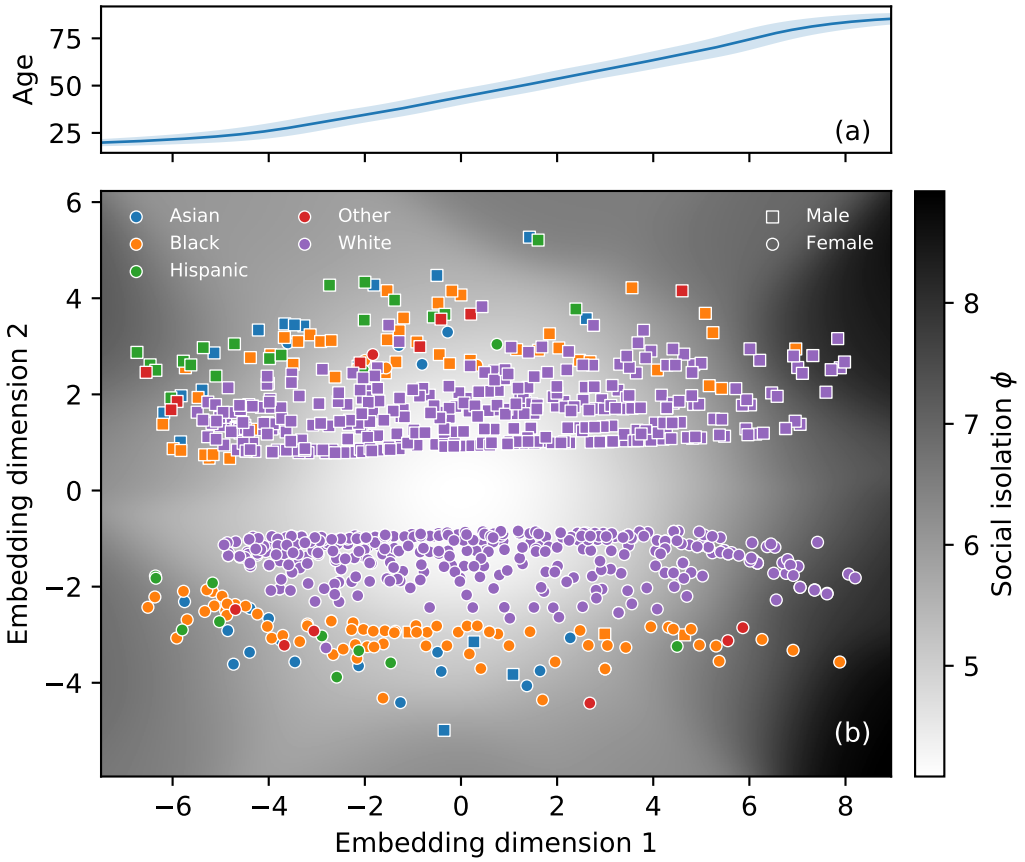


Figure 2: *A lower-dimensional embedding of the inter-node distances reveals an interpretable social space.* Panel (a) shows the mean and standard deviation of ages as a function of the first embedding dimension as a solid line and a shaded region, respectively. Panel (b) shows a scatter plot of respondents in a two-dimensional embedding space whose coordinates were obtained from the social separation φ using multidimensional scaling. The colour of a marker indicates the respondent's ethnicity and the shape indicates their sex. The heat map represents a smoothed estimate of the social isolation ϕ.

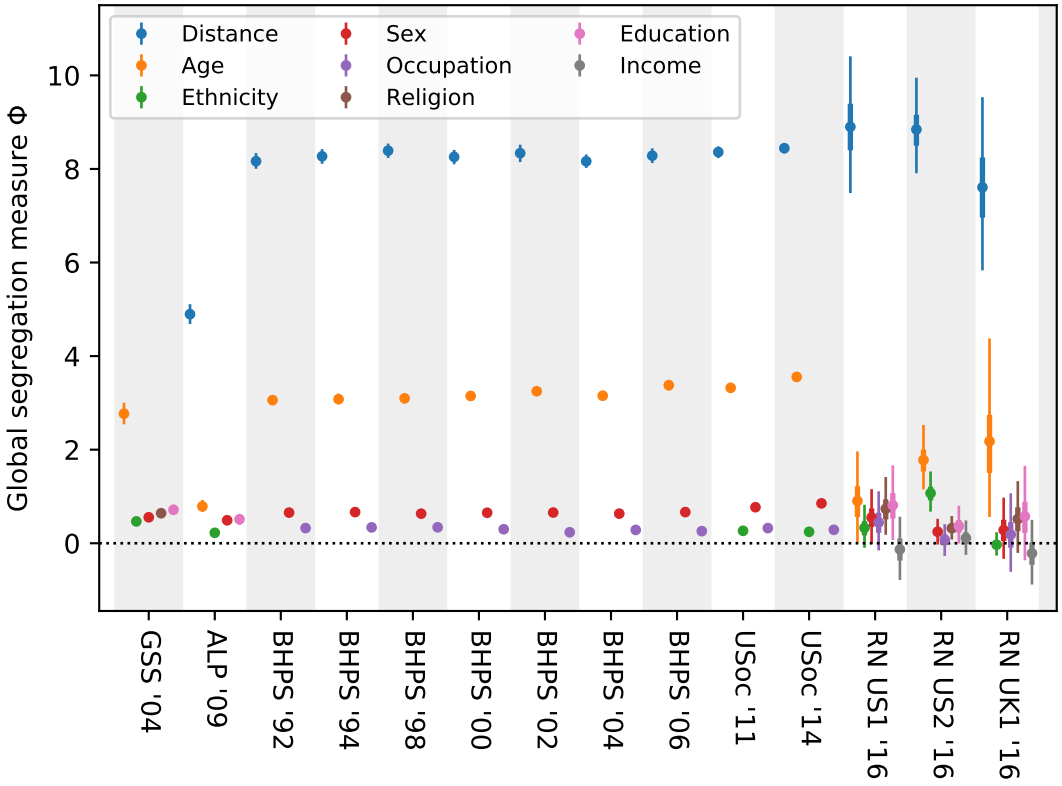


Figure 3: *Physical separation and age differences are the most important factors preventing integration of society for all datasets.* Markers represent the posterior mean of the contributions to social strain for each feature, thick error bars the interquartile range, and thin error bars the 95% posterior interval.

quantify segregation. Furthermore, Φ and its contributions in eq. (19) have an intuitive interpretation and universal scale, facilitating comparison across different datasets. In contrast, whilst standardising the features makes the comparison of kernel parameters easier [74], correlations amongst the features and skewed feature distributions can make them misleading [85]. For example, physical space has by far the most significant impact on how members of society connect with one another, as shown in fig. 3. The consistent, large contribution to the social strain due to spatial effects in the BHPS, USoc survey, and ALP survey are at odds with the varying magnitude of kernel parameters we inferred in sections 3.2 to 3.4. A possible explanation for the difference between kernel parameters and the social strain is that individuals adjust their preferences based on the local population density [24]. In other words, the number of "intervening opportunities" [26, 86] for making connections may be more important than the physical distance such that individuals in low-density areas are less discriminative and form longer-range ties [87].

The importance of physical space highlights that our suite of segregation measures does not distinguish between *choice homophily* and *opportunistic homophily* [88]. The former is a result of individuals having an active preference to connect with others who are alike, whereas the latter is a result of individuals being exposed to others who are similar to them. For example, residents of a neighbourhood that is largely homogeneous with respect to ethnicity may not have an active preference for connecting with others of the same ethnicity, but most friendships are likely to be homogeneous because there are fewer opportunities to connect with others of a different ethnicity. Opportunistic homophily is likely to be a large contributing factor to spatial homophily because individuals are less likely to encounter people who live far from them.

Age homophily places the second strongest restriction on how people connect with one another, and it is more than three times as restrictive as any other feature except physical space. The ALP survey is an exception because of the regression dilution [78] discussed in section 3.3. Age homophily is known to be particularly strong for friendship networks [1].

Homophily with respect to sex, ethnicity, education, religion, and occupation make similar, small contributions to the segregation of society. Importantly, the social strain captures the average contribution to social isolation: it can be small either because there is little homophily or because there is a large majority group, as evident from eq. (6). For example, almost 80% of respondents in the GSS identify as "White" and experience little social isolation due to their ethnicity, whereas minority groups experience more social isolation. On average, social isolation due to ethnicity is small.

4. Discussion

We considered a generative model for social networks embedded in Blau space, a space spanned by the demographic attributes of members of society. Investigating social distance in Blau space, we developed a family of segregation measures with a universal scale, facilitating comparison between datasets collected at different times or in different cultural contexts. Furthermore, the segregation measures are applicable to mixed

attribute types, have a natural reference point, and an intuitive interpretation. Based on eleven ego network datasets collected in the United Kingdom and United States, we inferred the connectivity kernel $\rho(x, y)$, i.e. the probability for an individual with demographic attributes x to connect with another with attributes y. Using the kernel, we compared segregation across different datasets along different demographic dimensions and found that physical distance and age have the most significant impact on how well society is connected.

Even though we did not expect the kernel parameters to be consistent across countries, time, or even different surveys, people connected with one another in a surprisingly similar fashion across the different datasets (the BHPS and USoc are longitudinal studies such that consistent parameter estimates are less surprising). Our observations, together with a study by Mossong et al. [89] finding that "mixing patterns [...] were remarkably similar across different European countries", suggest that a universal connectivity kernel for friendships may exist. To test this hypothesis, further surveys should be conducted in a unified fashion to minimise the effects of question wording and how the survey is administered [50]. In particular, such surveys should explore options to explicitly incentivise nominees to provide data about themselves [90]: seeds may not recall certain attributes, or nominees may deliberately portray themselves inaccurately [91]. Questions regarding ethnicity should allow respondents to provide multiple answers such that people with mixed ethnic backgrounds can express their identity. Rather than asking respondents about potentially sensitive information, such as income, proxy information that is more readily available—and potentially more informative of how individuals interact with society—could be collected [92]. Whenever possible, aggregation of attributes such as age into bins should be avoided because it limits the ability to infer kernel parameters [78], as we saw in section 3.3.

The connectivity kernel is an intuitive model of how people connect with one another, and it is able to reproduce some of the statistics of real social networks. For example, people in high-density regions of Blau space have been observed to have more connections [7]. However, exponential random graph models may be able to better capture the nature of social networks [93]. Furthermore, we have used a connectivity kernel that: (a) is symmetric and cannot identify whether there is a status order in society [19, 62], and (b) only depends on differences between individuals. For example, young men tend to have more social contacts than young women, and older women have more social contacts than older men [94]—an observation that cannot be captured by a kernel of the form we have considered. The connectivity kernel could be refined by adding the demographic attributes of the seeds and nominees as features, capturing sociability and popularity, respectively. Furthermore, it should be determined whether the number of "intervening opportunities" [86], absolute distance in Blau space, or a hybrid thereof are most predictive of tie probability. Ultimately, learning a connectivity kernel without a pre-specified parametric form should be considered [95]. But interpretable summary statistics, such as the family of segregation measures we considered here, are useful for comparisons across societies even for non-parametric kernels.

Because our approach is based on a generative model of social networks, it can generate synthetic social networks which can be studied without privacy risks [59]. Similarly,

the model can be used to impute missing attributes of individuals in social networks. In contrast to other methods, which generally only produce point estimates [10, 24], our approach yields a posterior distribution over demographic attributes that captures uncertainties.

For certain logistic connectivity kernels, the social separation is a metric for Blau space and allows us to quantify social distance in a principled fashion. We used the metric to evaluate the social distance amongst respondents to the GSS. Using a lower-dimensional embedding of the respondents, we explored Blau space, corroborating our findings that age has a profound impact on restricting friendship formation.

Whilst the social strain is able to summarise segregation at the level of an entire society, it is not suitable for studying the isolation experienced by minority groups because of its utilitarian foundations. Future work should consider other summary statistics of the social separation and isolation measures in the context of particular social questions. For example, using equal weights for different ethnicities in eq. (5), rather than weighting by their prevalence in the general population, would yield a measure that does not disadvantage minority groups. Unfortunately, equal weighting does not easily generalise for continuous attributes and further efforts are required.

References

[1] M. McPherson, L. Smith-Lovin and J. M. Cook. 'Birds of a Feather: Homophily in Social Networks'. In: *Annual Review of Sociology* 27 (2001), pp. 415–444. DOI: 10.1146/annurev.soc.27.1.415.

[2] B. Golub and M. O. Jackson. 'How Homophily Affects the Speed of Learning and Best-Response Dynamics'. In: *The Quarterly Journal of Economics* 127.3 (2012), pp. 1287–1338. DOI: 10.1093/qje/qjs021.

[3] A. Boutyline and R. Willer. 'The Social Structure of Political Echo Chambers: Variation in Ideological Homophily in Online Networks'. In: *Political Psychology* 38.3 (2017), pp. 551–569. DOI: 10.1111/pops.12337.

[4] E. Bakshy, S. Messing and L. A. Adamic. 'Exposure to ideologically diverse news and opinion on Facebook'. In: *Science* 348.6239 (2015), pp. 1130–1132. DOI: 10.1126/science.aaa1160.

[5] P. M. DeMarzo, D. Vayanos and J. Zwiebel. 'Persuasion Bias, Social Influence, and Unidimensional Opinions'. In: *The Quarterly Journal of Economics* 118.3 (2003), pp. 909–968. DOI: 10.1162/00335530360698469.

[6] M. Salathé and S. Bonhoeffer. 'The effect of opinion clustering on disease outbreaks'. In: *Journal of The Royal Society Interface* 5.29 (2008), pp. 1505–1508. DOI: 10.1098/rsif.2008.0271.

[7] S. Currarini, M. O. Jackson and P. Pin. 'An economic model of friendship: homophily, minorities, and segregation'. In: *Econometrica* 77.4 (2009), pp. 1003–1045. DOI: 10.3982/ECTA7528.

[8] J. R. Hipp and A. J. Perrin. 'The Simultaneous Effect of Social Distance and Physical Distance on the Formation of Neighborhood Ties'. In: *City & Community* 8.1 (2009), pp. 5–25. DOI: 10.1111/j.1540-6040.2009.01267.x.

[9] M. McPherson, L. Smith-Lovin and M. E. Brashears. 'Social Isolation in America: Changes in Core Discussion Networks over Two Decades'. In: *American Sociological Review* 71.3 (2006), pp. 353–375. DOI: 10.1177/000312240607100301.

[10] Y. Wang, H. Zang and M. Faloutsos. 'Inferring cellular user demographic information using homophily on call graphs'. In: *IEEE Conference (Computer Communications Workshops)*. 2013, pp. 211–216. DOI: 10.1109/INFCOMW.2013.6562897.

[11] Y. Leo et al. 'Socioeconomic correlations and stratification in social-communication networks'. In: *Journal of The Royal Society Interface* 13.125 (2016), p. 20160598. DOI: 10.1098/rsif.2016.0598.

[12] P. Blau and J. Schwartz. *Crosscutting Social Circles: Testing a Macrostructual Theory of Integroup Relations*. Routledge, 1984.

[13] J. Chang et al. 'ePluribus: ethnicity on social networks'. In: *ICWSM*. 2010.

[14] P. V. Marsden. 'Homogeneity in confiding relations'. In: *Social Networks* 10.1 (1988), pp. 57–76. DOI: 10.1016/0378-8733(88)90010-X.

[15] J. A. Smith, M. McPherson and L. Smith-Lovin. 'Social Distance in the United States: Sex, Race, Religion, Age, and Education Homophily among Confidants, 1985 to 2004'. In: *American Sociological Review* 79.3 (2014), pp. 432–456. DOI: 10.1177/0003122414531776.

[16] J. Stehlé et al. 'High-Resolution Measurements of Face-to-Face Contact Patterns in a Primary School'. In: *PLOS ONE* 6.8 (2011), pp. 1–13. DOI: 10.1371/journal.pone.0023176.

[17] J. Blumenstock and L. Fratamico. 'Social and spatial ethnic segregation: A framework for analyzing segregation with large-scale spatial network data'. In: *4th Annual Symposium on Computing for Development*. 2013. DOI: 10.1145/2537052.2537061.

[18] M. A. Johnson. 'Variables Associated with Friendship in an Adult Population'. In: *The Journal of Social Psychology* 129.3 (1989), pp. 379–390. DOI: 10.1080/00224545.1989.9712054.

[19] T. W. Chan and J. H. Goldthorpe. 'Is There a Status Order in Contemporary British Society? Evidence from the Occupational Structure of Friendship'. In: *European Sociological Review* 20.5 (2004), pp. 383–401. DOI: 10.1093/esr/jch033.

[20] L. Platt. 'Muslims in Britain: making social and political space'. In: ed. by W. Ahmad and Z. Sardar. Routledge, 2012. Chap. Exploring social spaces of Muslims, pp. 53–83.

[21] M. Kalmijn and J. K. Vermunt. 'Homogeneity of social networks by age and marital status: A multilevel analysis of ego-centered networks'. In: *Social Networks* 29.1 (2007), pp. 25–43. DOI: 10.1016/j.socnet.2005.11.008.

[22] R. Lambiotte et al. 'Geographical dispersal of mobile communication networks'. In: *Physica A* 387.21 (2008), pp. 5317–5325. DOI: 10.1016/j.physa.2008.05.014.

[23] P. Expert et al. 'Uncovering space-independent communities in spatial networks'. In: *PNAS* 108.19 (2011), pp. 7663–7668. DOI: 10.1073/pnas.1018962108.

[24] L. Backstrom, E. Sun and C. Marlow. 'Find me if you can: improving geographical prediction with social and spatial proximity'. In: *WWW*. 2010, pp. 61–70. DOI: 10.1145/1772690.1772698.

[25] S. Scellato et al. 'Socio-spatial properties of online location-based social networks.' In: *ICWSM*. 2011, pp. 329–336.

[26] J. Illenberger, K. Nagel and G. Flötteröd. 'The Role of Spatial Interaction in Social Networks'. In: *Networks and Spatial Economics* 13.3 (2013), pp. 255–282. DOI: 10.1007/s11067-012-9180-4.

[27] A. Rodriguez-Moral and M. Vorsatz. 'Complex Networks and Dynamics'. In: ed. by P. Commendatore et al. Springer, 2016. Chap. An Overview of the Measurement of Segregation: Classical Approaches and Social Network Analysis, pp. 93–119. DOI: 10.1007/978-3-319-40803-3_5.

[28] M. Bojanowski and R. Corten. 'Measuring segregation in social networks'. In: *Social Networks* 39 (2014), pp. 14–32. DOI: 10.1016/j.socnet.2014.04.001.

[29] G. Orfield and E. Frankenberg. *Brown at 60: Great Progress, a Long Retreat and an Uncertain Future*. Tech. rep. Civil Rights Project, 2014.

[30] P. A. Popielarz. '(In)voluntary Association: A Multilevel Analysis of Gender Segregation in Voluntary Organizations'. In: *Gender & Society* 13.2 (1999), pp. 234–250. DOI: 10.1177/089124399013002005.

[31] M. Charles and D. B. Grusky. 'Models for Describing the Underlying Structure of Sex Segregation'. In: *American Journal of Sociology* 100.4 (1995), pp. 931–971. DOI: 10.1086/230605.

[32] S. F. Reardon and D. O'Sullivan. 'Measures of Spatial Segregation'. In: *Sociological Methodology* 34.1 (2004), pp. 121–162. DOI: 10.1111/j.0081-1750.2004.00150.x.

[33] S. F. Reardon. *Measures of Income Segregation*. Tech. rep. Stanford, 2011.

[34] J. Moody. 'Race, School Integration, and Friendship Segregation in America'. In: *American Journal of Sociology* 107.3 (2001), pp. 679–716. DOI: 10.1086/338954.

[35] M. E. J. Newman. 'Mixing patterns in networks'. In: *Phys. Rev. E* 67.2 (2003), p. 026126. DOI: 10.1103/PhysRevE.67.026126.

[36] D. Lam Morgan. 'A Spatial Econometric Approach To The Study Of Social Influence'. PhD thesis. University of Texas at Austin, 2012.

[37] M. Kim and J. Leskovec. 'Multiplicative Attribute Graph Model of Real-World Networks'. In: *Internet Mathematics* 8.1–2 (2012), pp. 113–160. DOI: 10.1080/15427951.2012.625257.

[38] C. T. Butts et al. 'Geographical variability and network structure'. In: *Social Networks* 34.1 (2012), pp. 82–100. DOI: 10.1016/j.socnet.2011.08.003.

[39] S. A. Golder and M. W. Macy. 'Digital Footprints: Opportunities and Challenges for Online Social Research'. In: *Annual Review of Sociology* 40.1 (2014), pp. 129–152. DOI: 10.1146/annurev-soc-071913-043145.

[40] J. Blumenstock, G. Cadamuro and R. On. 'Predicting poverty and wealth from mobile phone metadata'. In: *Science* 350.6264 (2015), pp. 1073–1076. DOI: 10.1126/science.aac4420.

[41] S. Luo et al. 'Inferring personal economic status from social network location'. In: 8 (2017), p. 15227. DOI: 10.1038/ncomms15227.

[42] Y. Wang and M. Kosinski. 'Deep neural networks are more accurate than humans at detecting sexual orientation from facial images.' In: *Open Science Framework* (2017), zn79k.

[43] M. Kosinski, D. Stillwell and T. Graepel. 'Private traits and attributes are predictable from digital records of human behavior'. In: *PNAS* 110.15 (2013), pp. 5802–5805. DOI: 10.1073/pnas.1218772110.

[44] L. Backstrom, C. Dwork and J. Kleinberg. 'Wherefore Art Thou R3579X?: Anonymized Social Networks, Hidden Patterns, and Structural Steganography'. In: *Communications of the ACM* 54.12 (2011), pp. 133–141. DOI: 10.1145/2043174.2043199.

[45] A. Narayanan and V. Shmatikov. 'Robust de-anonymization of Large Sparse Datasets'. In: *IEEE Symposium on Security and Privacy*. 2008, pp. 111–125. DOI: 10.1109/SP.2008.33.

[46] R. R. Huckfeldt. 'Social Contexts, Social Networks, and Urban Neighborhoods: Environmental Constraints on Friendship Choice'. In: *American Journal of Sociology* 89.3 (1983), pp. 651–669. DOI: 10.1086/227908.

[47] P. V. Marsden. 'Core Discussion Networks of Americans'. In: *American Sociological Review* 52.1 (1987), pp. 122–131. DOI: 10.2307/2095397.

[48] A. Banerjee et al. 'The Diffusion of Microfinance'. In: *Science* 341.6144 (2013), p. 1236498. DOI: 10.1126/science.1236498.

[49] A. Marin. 'Are respondents more likely to list alters with certain characteristics? Implications for name generator data'. In: *Social Networks* 26.4 (2004), pp. 289–307. DOI: 10.1016/j.socnet.2004.06.001.

[50] D. E. Eagle and R. J. Proeschold-Bell. 'Methodological considerations in the use of name generators and interpreters'. In: *Social Networks* 40 (2015), pp. 75–83. DOI: 10.1016/j.socnet.2014.07.005.

[51] W. P. Eveland Jr., O. Appiah and P. A. Beck. 'Americans are more exposed to difference than we think: Capturing hidden exposure to political and racial difference'. In: *Social Networks* 52 (2017), pp. 192–200. DOI: 10.1016/j.socnet.2017.08.002.

[52] T. A. Snijders. 'Statistical Models for Social Networks'. In: *Annual Review of Sociology* 37.1 (2011), pp. 131–153. DOI: 10.1146/annurev.soc.012809.102709.

[53] P. D. Hoff, A. E. Raftery and M. S. Handcock. 'Latent Space Approaches to Social Network Analysis'. In: *Journal of the American Statistical Association* 97.460 (2002), pp. 1090–1098. DOI: 10.1198/016214502388618906.

[54] M. S. Handcock, A. E. Raftery and J. M. Tantrum. 'Model-based clustering for social networks'. In: *Journal of the Royal Statistical Society A* 170.2 (2007), pp. 301–354. DOI: 10.1111/j.1467-985X.2007.00471.x.

[55] A. E. Raftery et al. 'Fast inference for the latent space network model using a case-control approximate likelihood'. In: *Journal of Computational and Graphical Statistics* 21.4 (2012), pp. 9010–919. DOI: 10.1080/10618600.2012.679240.

[56] J. J. Pfeiffer III et al. 'Attributed Graph Models: Modeling Network Structure with Correlated Attributes'. In: *WWW*. 2014, pp. 831–842. DOI: 10.1145/2566486.2567993.

[57] S. Lieberman and J. K. Alt. 'Advances in Social Computing'. In: ed. by S.-K. Chai, J. J. Salerno and P. L. Mabry. Springer, 2010. Chap. Developing Social Networks for Artificial Societies from Survey Data, pp. 159–168. DOI: 10.1007/978-3-642-12079-4_21.

[58] S. Lieberman. 'Extensible software for whole of society modeling: framework and preliminary results'. In: *Simulation* 88.5 (2012), pp. 557–564. DOI: 10.1177/0037549711404918.

[59] D. F. Nettleton. 'A synthetic data generator for online social network graphs'. In: *Social Network Analysis and Mining* 6.1 (2016), p. 44. DOI: 10.1007/s13278-016-0352-y.

[60] A. V. Sathanur et al. 'When Labels Fall Short: Property Graph Simulation via Blending of Network Structure and Vertex Attributes'. In: *CIKM*. 2017, pp. 2287–2290. DOI: 10.1145/3132847.3133065.

[61] P. D. Hoff. 'Multiplicative latent factor models for description and prediction of social networks'. In: *Computational and Mathematical Organization Theory* 15.4 (2008), pp. 261–272. DOI: 10.1007/s10588-008-9040-4.

[62] B. Ball and M. Newman. 'Friendship networks and social status'. In: *Network Science* 1.1 (2013), pp. 16–30. DOI: 10.1017/nws.2012.4.

[63] S. E. Fienberg. 'A Brief History of Statistical Models for Network Analysis and Open Challenges'. In: *Journal of Computational and Graphical Statistics* 21.4 (2012), pp. 825–839. DOI: 10.1080/10618600.2012.738106.

[64] L. Barnett, E. Di Paolo and S. Bullock. 'Spatially embedded random networks'. In: *Phys. Rev. E* 76 (2007), p. 056115. DOI: 10.1103/PhysRevE.76.056115.

[65] F. Caron and E. B. Fox. 'Sparse graphs using exchangeable random measures'. In: *Journal of the Royal Statistical Society B* 79.5 (2017), pp. 1295–1366. DOI: 10.1111/rssb.12233.

[66] L. C. Freeman. 'Segregation in Social Networks'. In: *Sociological Methods & Research* 6.4 (1978), pp. 411–429. DOI: 10.1177/004912417800600401.

[67] D. Krackhardt and R. N. Stern. 'Informal Networks and Organizational Crises: An Experimental Simulation'. In: *Social Psychology Quarterly* 51.2 (1988), pp. 123–140. DOI: 10.2307/2786835.

[68] W. A. Wilson. 'On Semi-Metric Spaces'. In: *American Journal of Mathematics* 53.2 (1931), pp. 361–373. DOI: 10.2307/2370790.

[69] P. M. Blau. 'A Macrosociological Theory of Social Structure'. In: *American Journal of Sociology* 83.1 (1977), pp. 26–54. DOI: 10.1086/226505.

[70] G. King and L. Zeng. 'Logistic Regression in Rare Events Data'. In: *Political Analysis* 9.2 (2001), pp. 137–163. DOI: 10.1093/oxfordjournals.pan.a004868.

[71] D. B. Rubin. 'Inference and missing data'. In: *Biometrika* 63.3 (1976), pp. 581–592. DOI: 10.1093/biomet/63.3.581.

[72] T. D. Pigott. 'A Review of Methods for Missing Data'. In: *Educational Research and Evaluation* 7.4 (2001), pp. 353–383. DOI: 10.1076/edre.7.4.353.8937.

[73] A. Gelman et al. *Bayesian Data Analysis*. 3rd ed. Chapman and Hall/CRC, 2013.

[74] A. Gelman. 'Scaling regression inputs by dividing by two standard deviations'. In: *Statistics in Medicine* 27.15 (2008), pp. 2865–2873. DOI: 10.1002/sim.3107.

[75] A. Gelman et al. 'A weakly informative default prior distribution for logistic and other regression models'. In: *Ann. Appl. Stat.* 2.4 (2008), pp. 1360–1383. DOI: 10.1214/08-AOAS191.

[76] W. K. Hastings. 'Monte Carlo sampling methods using Markov chains and their applications'. In: *Biometrika* (1970), pp. 97–109. DOI: 10.1093/biomet/57.1.97.

[77] K. Mihaly. *American Life Panel: Well-being 86 questionnaire*. Tech. rep. RAND corporation, 2009.

[78] J. A. Hutcheon, A. Chiolero and J. A. Hanley. 'Random measurement error and regression dilution bias'. In: *BMJ* 340 (2010), pp. 1402–1406. DOI: 10.1136/bmj.c2289.

[79] Institute for Social and Economic Research. *British Household Panel Survey Questionnaire Wave 10*. 2000. URL: https://www.iser.essex.ac.uk/bhps/documentation/pdf_versions/questionnaires/bhpsw10q.pdf.

[80] Institute for Social and Economic Research. *UK Household Longitudinal Study Mainstage Questionnaire Wave 3*. 2017. URL: https://www.understandingsociety.ac.uk/documentation/mainstage/questionnaire/questionnaire-documents/mainstage/wave-3/Understanding_Society_Wave_3_Questionnaire_v03.pdf.

[81] T. Hoffmann and V. Nolan. *Understanding Society User Support: Unusual age distribution after conditioning on wJBSTATT in wave R*. 2016. URL: https://www.understandingsociety.ac.uk/support/issues/687.

[82] T. Hoffmann and S. Auty. *Understanding Society User Support: Unexpectedly strong gender homophily in Understanding Society compared with the BHPS*. 2017. URL: https://www.understandingsociety.ac.uk/support/issues/869.

[83] I. Borg and P. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer, 1996. DOI: 10.1007/0-387-28981-X.

[84] T. Hastie, R. Tibshirani and J. Friedman. *The elements of statistical learning: data mining, inference and prediction*. 2nd ed. Springer, 2009. DOI: 10.1007/978-0-387-84858-7.

[85] S. Greenland, J. J. Schlesselman and M. H. Criqui. 'The Fallacy Of Employing Standardized Regression Coefficients And Correlations As Measures Of Effect'. In: *American Journal of Epidemiology* 123.2 (1986), pp. 203–208. DOI: 10.1093/oxfordjournals.aje.a114229.

[86] S. A. Stouffer. 'Intervening Opportunities: A Theory Relating Mobility and Distance'. In: *American Sociological Review* 5.6 (1940), pp. 845–867. DOI: 10.2307/2084520.

[87] D. Liben-Nowell et al. 'Geographic routing in social networks'. In: *PNAS* 102.33 (2005), pp. 11623–11628. DOI: 10.1073/pnas.0503018102.

[88] S. Franz, M. Marsili and P. Pin. 'Observed Choices And Underlying Opportunities'. In: *Science and Culture* 76.9–10 (2010), pp. 471–476.

[89] J. Mossong et al. 'Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases'. In: *PLOS Medicine* 5.3 (2008), e74. DOI: 10.1371/journal.pmed.0050074.

[90] P. Biernacki and D. Waldorf. 'Snowball Sampling: Problems and Techniques of Chain Referral Sampling'. In: *Sociological Methods & Research* 10.2 (1981), pp. 141–163. DOI: 10.1177/004912418101000205.

[91] E. Bruch, F. Feinberg and K. Y. Lee. 'Extracting multistage screening rules from online dating activity data'. In: *PNAS* 113.38 (2016), pp. 10530–10535. DOI: 10.1073/pnas.1522494113.

[92] J. Y. T. Po et al. *Estimating Household Permanent Income from Ownership of Physical Assets*. Working paper 97. Harvard, 2012.

[93] A. Wimmer and K. Lewis. 'Beyond and Below Racial Homophily: ERG Models of a Friendship Network Documented on Facebook'. In: *American Journal of Sociology* 116.2 (2010), pp. 583–642. DOI: 10.1086/653658.

[94] K. Bhattacharya et al. 'Sex differences in social focus across the life cycle in humans'. In: *Open Science* 3.4 (2016), p. 160097. DOI: 10.1098/rsos.160097.

[95] M. Frölich. 'Non-parametric regression for binary dependent variables'. In: *Econometrics Journal* 9.3 (2006), pp. 511–540. DOI: 10.1111/j.1368-423X.2006.00196.x.

[96] L. Kish. 'Weighting for unequal P_i'. In: *Journal of Official Statistics* 8.2 (1992), pp. 183–200.

[97] A. Gelman. 'Struggles with survey weighting and regression modeling'. In: *Statistical Science* 22.2 (2007), pp. 153–164. DOI: 10.1214/088342306000000691.

[98] D. Pfeffermann. 'The use of sampling weights for survey data analysis'. In: *Statistical Methods in Medical Research* 5.3 (1996), pp. 239–261. DOI: 10.1177/096228029600500303.

[99] Department for Communities and Local Government. *Lower layer super output areas*. URL: http://opendatacommunities.org/data/lower-layer-super-output-areas (visited on 18/01/2018).

[100] J. A. Smith. 'A Social Space Approach to Testing Complex Hypotheses: The Case of Hispanic Marriage Patterns in the United States'. In: *Socius* 3 (2017), p. 2378023117739176. DOI: 10.1177/2378023117739176.

A. Evaluation of the observed-data log-likelihood

A.1. Weighting to account for non-uniform inclusion probabilities

Seeds are often not included in the survey uniformly at random, and weights are traditionally used to compensate for the potentially biased selection of respondents [96]. Including weights in Bayesian analyses is generally difficult [97], and, in principle, we should model the data collection process explicitly [73, chapter 8]. Unfortunately, modelling the data collection process is non-trivial, and we use a weighted pseudo-likelihood instead [98]. In particular, the observed-data log-likelihood from eq. (18) becomes

$$L = \sum_{(i,j):I_{ij}=1} w_j \left\{A_{ij} + (1 - A_{ij})w_i\right\} \times \left\{A_{ij} \log \rho_{ij} + (1 - A_{ij}) \log(1 - \rho_{ij}) - \log\left[r(0)(1 - \rho_{ij}) + r(1)\rho_{ij}\right]\right\}, \quad \text{(A.1)}$$

where w_j is the weight associated with seed j. We clip all weights exceeding the 95^{th} percentile of the empirical weight distribution and normalise them such that $\sum_{j \in U} w_j = |U|$. Censoring the weights, also known as Winsorisation, limits the variance induced by attributing variable importance to different observations at the expense of introducing a small bias [96].

A.2. Numerical stability

The evaluation of the observed-data log-likelihood may suffer from numerical instabilities, especially when the connectivity kernel is small. We can mitigate such instabilities for logistic connectiviy kernels, i.e.

$$\rho(x, y, \theta) = \sigma(\theta^\intercal f(x, y)), \tag{A.2}$$

$$\text{where } \sigma(\xi) = \frac{1}{1 + \exp(-\xi)} \tag{A.3}$$

is the logistic function. In particular, note that $1 - \sigma(\xi) = \sigma(-\xi)$ and $\log \sigma(\xi) = -\operatorname{log1p} \exp(-\xi)$, where $\operatorname{log1p}(\xi) = \log(1 + \xi)$ is a numerically stable implementation—even for $|\xi| \ll 1$. Substituting into eq. (18) yields

$$\begin{aligned}\log P(A|f, \theta, I = 1) = &- \sum_{(i,j):I_{ij}=1} A_{ij} \operatorname{log1p} \exp\left(-\theta^\intercal f_{ij}\right) + (1 - A_{ij}) \operatorname{log1p} \exp\left(\theta^\intercal f_{ij}\right) \\ &+ \operatorname{logsumexp}\left[\log r(0) - \operatorname{log1p} \exp\left(\theta^\intercal f_{ij}\right) +, \log r(1) - \operatorname{log1p} \exp\left(-\theta^\intercal f_{ij}\right)\right], \end{aligned} \tag{A.4}$$

where $\operatorname{logsumexp}(x_1, \ldots, x_k) = \log \sum_{i=1}^{k} \exp(x_k)$ is a numerically stable implementation.

B. Coding of demographic attributes and feature maps

In the BHPS and USoc surveys, distance was coded as an ordinal variable: less than one mile, less than five miles, less than fifty miles, and more than fifty miles. We rely on having complete information about seeds to evaluate the control features in eq. (7). But data on the residential location of seeds is not made available to protect their privacy. Fortunately, we can sample the home locations of respondents[1] using population estimates and the geographic boundaries of lower layer super output areas (LSOAs). LSOAs are census reporting areas and have a few thousand inhabitants each [99]. We approximate the distribution of distances between residents of the UK using rejection sampling: first, choose a LSOA with probability proportional to the number of residents. Second, choose one of the polygons associated with the LSOA with probability proportional to the area of the polygon (LSOAs are not necessarily contiguous). Third, sample points uniformly inside the bounding box of the polygon until a point inside the polygon is sampled. The last two steps assume uniform population densities within each LSOA, which is unlikely to be problematic as they are small areas. Having sampled the residential location of two respondents, we calculate the distance between respondents and cast to the same ordinal scale as reported for nominees. USoc furthermore distinguishes between friends living more than fifty miles apart but within the UK and friends outside the UK. We discard the latter (3.6% and 2.8% of all friends in waves C and F of USoc) because it is difficult to define an appropriate control population. For the BHPS, we implicitly assume that all friends are resident in the UK.

[1] Sampling home locations cannot reproduce any correlation between home location and other demographic attributes.

In USoc, respondents could identify with mixed ethnicities, and we coded such responses as a mixed membership. For example, a respondent who indicated "mixed Asian and White" would belong to both White and Asian ethnicities. To measure how different two people are in terms of ethnicity, we define the feature map

$$f_{\text{ethnicity}}(x_i, x_j) = \frac{1}{2} \sum_{l \in E} |x_{il} - x_{jl}| \,,$$

where E is the set of attributes encoding ethnic identity, and ethnicity memberships are normalised such that $\sum_{l \in E} x_{jl} = 1$ for all j. For example, $f_{\text{ethnicity}}(x_i, x_j) = 1$ for two people i and j one of which identifies as white and the other as black. For a person i identifying as black and another person j identifying as mixed black and Asian, $f_{\text{ethnicity}}(x_i, x_j) = 0.5$.

Variable	Seed coding	Nominee coding	$f(x, y)$
Bias term	.....		1
Age	Age in years		$\lvert x - y \rvert$
Sex	(a) Male, (b) Female		$x \neq y$
Ethnicity	(a) {Asian Indian, Chinese, Filipino, Japanese, Korean, Vietnamese, Other Asian} (b) Black, (c) Hispanic, (d) White, (e) {American Indian or Alaska Native, Native Hawaiian, Guamanian or Chamorro, Samoan, Other Pacific Islander, Other}	(a) Asian, (b) Black, (c) Hispanic, (d) White, (e) Other	$x \neq y$
Religion	(a) Protestant, (b) Catholic, (c) Jewish, (d) None, (e) {Other, Buddhism, Hinduism, Islam, Orthodox, Christian, Native American, Nondenominational}	(a) Protestant, (b) Catholic, (c) Jewish, (d) None, (e) Other	$x \neq y$
Education	(1) 1–6 years, (2) 7–12 years without high school diploma, (3) exactly 12 years with high school diploma, (4) $>$ 12 years without degree, (5) Associate degree, (6) Bachelor's degree, (7) Professional or graduate degree	(1) 1–6 years, (2) 7–12 years, (3) High school graduate, (4) Some college, (5) Associate degree, (6) Bachelor's degree, (7) Professional or graduate degree	$\lvert x - y \rvert$

Table B.1: *Coding of the demographic variables for the General Social Survey together with the feature maps for each variable.* Seeds were provided with 16 options to choose from for their own ethnicity but only five options for their nominees. We attempt to unify the educational coding by combining the number of years of education and formal qualifications of the seeds to approximate the coding of nominees. The bias term in the first row of the table controls the overall edge density.

Dataset	Seeds	Dropped seeds	Nominees	Dropped nominees
GSS 2004	2,785	27 (1.0%)	869	164 (15.9%)
ALP 2009	2,511	0 (0.0%)	895	171 (16.0%)
BHPS 1992	9,427	32 (0.3%)	19,126	601 (3.0%)
BHPS 1994	9,019	38 (0.4%)	17,609	481 (2.7%)
BHPS 1998	8,864	30 (0.3%)	17,309	459 (2.6%)
BHPS 2000	8,603	23 (0.3%)	16,742	441 (2.6%)
BHPS 2002	8,277	18 (0.2%)	16,276	496 (3.0%)
BHPS 2004	7,937	31 (0.4%)	15,430	455 (2.9%)
BHPS 2006	7,785	19 (0.2%)	14,836	383 (2.5%)
USoc 2011	37,718	505 (1.3%)	77,163	1,062 (1.4%)
USoc 2014	30,163	325 (1.1%)	64,630	829 (1.3%)

Table B.2: *Number of retained seeds and nominees for each dataset together with the number of individuals who have been excluded from the analysis because one or more of their demographic attributes were missing.*

Variable	Seed coding	Nominee coding	$f(x, y)$
Bias term	.....		1
Age	Age in years recoded to match the nominee coding	Age brackets in years: (1) 0–20, (2) 21–35, (3) 36–50, (4) 51–65, (5) 66–80, (6) > 80	$\lvert x - y \rvert$
Sex	(a) Male, (b) Female		$x \neq y$
Ethnicity	(a) White or Caucasian, (b) Black or African American, (c) American Indian or Alaska Native, (d) Asian or Pacific Islander, (e) Hispanic (see below) (f) Other		$x \neq y$
Hispanic	(a) Yes, (b) No; ethnicity is coded as "Hispanic" if response is affirmative.....		
Education	The seed coding is more refined but can be reduced to the nominee coding: (1) Less than 9^{th} grade, (2) 9^{th}–12^{th} grade without diploma, (3) High school graduate, (4) Some college, (5) Associate degree, (6) Bachelor's degree, (7) Master's degree, (8) Professional degree or doctorate.....		$\lvert x - y \rvert$
State	One of 52 states and Washington DC and Puerto Rico.....		$x \neq y$

Table B.3: *Coding of the demographic variables for the American Life Panel together with the feature maps for each variable.* We aggregate the ages and educational attainments of seeds to match the coarser coding of nominees, as shown in table B.3. The joint effect of ethnic differences and whether people identify as Hispanic is still unclear [100]; for consistency with the GSS, we code the ethnicity of respondents as "Hispanic" if they consider themselves to be Hispanic or Latino irrespective of their reported ethnicity. In fact, 46% of respondents who identified as Hispanic selected "other" as their ethnicity, compared with $< 1\%$ for respondents who did not identify as Hispanic.

Variable	Seed coding	Nominee coding	$f(x, y)$
Bias term	.....		1
Age	Age in years.....		$\lvert x - y \rvert$
Sex	(a) Male, (b) Female.....		$x \neq y$
Occupation	(a) {Self-employed, employed, maternity leave, unpaid worker in family business[a]}, (b) {Unemployed, disabled}, (c) {Full-time student, government training scheme}, (d) Family care, (e) Retired	(a) {Full-time employed, part-time employed}, (b) Unemployed, (c) Full-time education, (d) Full-time housework, (e) Retired	$x \neq y$
Distance	Only applicable to the seed-nominee pair	(1) $<$ 1 mile, (2) $<$ 5 miles, (3) $<$ 50 miles, (4) $\geq$ 50 miles but still in the UK	[b]
Ethnicity[a]	Independent binary choices: White, Asian, Black, Other.....		[c]

[a]Only available in Understanding Society.
[b]We use the ordinal distance measure reported in the survey as a regression feature and generate control features using Monte Carlo simulation, as discussed in section 3.4.
[c]See section 3.4 for a detailed description of the feature map.

Table B.4: *Coding of the demographic variables for the British Household Panel Survey and Understanding Society together with the feature maps for each variable.* Sex and age have identical coding for seeds and nominees. We aggregate the detailed occupational coding of seeds to match the coding of nominees. In particular, we code women on maternity leave as employed because their occupational status is only temporary, and we code disabled individuals as "not employed" because they are unlikely to have the same social opportunities as people in employment.