Inference of a universal social scale and segregation measures using social connectivity kernels

Hoffmann, Till and Jones, Nick S.

Imperial College London

28 October 2020

Inference of a universal social scale and segregation measures using social connectivity kernels

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, allowing comparison within and between societies, remains elusive.

We use the connectivity kernel of conditionally-independent edge models to develop a family of segregation statistics with desirable properties: they offer an intuitive and universal characteristic scale on social space (facilitating comparison across datasets and societies), 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 show that the segregation statistics can induce a metric on Blau space (a space spanned by the attributes of the members of society) and provide maps of two societies.

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

Peter Blau proposed that individuals connect with one another based on their positions in a high-dimensional space [1], e.g. a space spanned by demographic attributes. With

an accrual of large-scale survey data, we now have access to demographic and relational information internationally, for whole societies, and over time [2, 3, 4]. Despite this wealth of data, we lack two important quantities: (a) a natural notion of distance in this space, allowing us to determine how far apart individuals are in society, and (b) a universal characteristic scale, allowing the distance between pairs of individuals in one society to be compared to another society with different social dimensions. We will argue that the probability of forming a friendship, intimately related to work on homophily by Blau and his successors [2], offers both a universal characteristicscale and a notion of distance.

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 [2]. 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 [5]. For example, users of online social networks, such as Facebook and Twitter, tend to connect with others who hold similar political views [6]. They are more likely to be exposed to information that confirms rather than challenges their beliefs [7]. An "echo chamber" effect ensues, leading to polarised opinions [8]. 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 [9]—even if the vaccination rate is above herd immunity levels on average.

Homophily can be observed in friendships [10, 11], networks of discussion partners [3], communication networks [12, 13], marital ties [14], and online social networks [15]. Relationships are homogeneous with respect to a wide range of attributes, including age [16, 17], sex [17], ethnicity [15, 18, 10], education [3, 17, 19], occupation [20], income [13, 12, 19], religion [21], parental [19] and marital status [22], political ideology [7, 6], and geographical location [23, 24, 25, 26, 27].

Segregation statistics, also referred to as segregation measures, are often used to quantify homophily [28, 29]. Many approaches are based on co-presence in organisational units such as schools [30], voluntary associations [31], occupations [32], or census tracts [33], and we refer to them as *organisational* statistics. Typically, they compare how the distribution of demographic attributes within organisational units differs from the distribution of attributes in the general population. Whilst organisational statistics are applicable whenever data can be stratified according to a variable of interest, they cannot capture segregation at smaller scales than the strata [18]. 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 [10, 34]. Organisational statistics cannot capture such social segregation.

Social statistics of segregation, such as the assortativity coefficient [35], overcome these limitations by explicitly considering the interactions amongst individuals [18], but have their own difficulties: first, they usually rely on the existence of discrete groups, such as sex, ethnicity, or religion [29], and they are not applicable to continuous attributes, such as age or income. Attributes are often discretised [36, 22, 37], but the boundaries between

categories always suffer from some degree of arbitrariness [33]. Second, segregation for multiple attributes can be quantified independently, but Pelechrinis and Wei [38], who consider a multivariate generalisation of the assortativity coefficient, note that "a formal metric that is generally applicable" remains elusive. Furthermore, social statistics 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 [39, 18]: 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 [40], and seemingly private traits such as socioeconomic status [41, 42], sexual orientation [43], age, gender, and political ideology can be inferred [44]. Unfortunately, network features are often used to predict demographic attributes [12, 41, 42, 44], 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 [40]. Individuals can be identified in anonymised social networks [45, 46], and augmenting the network data with demographic information would make re-identification even easier.

However, censuses and large-scale surveys collect comprehensive demographic information from respondents but usually lack data about their associates. Fortunately, some surveys have included questions about respondents' friends [47, 19], discussion partners [48, 3], or support networks [22, 49]. The questions used to elicit social ties provide an imperfect observation of the immediate neighbourhood of respondents [50, 51, 52].

Building on the successes of conditionally-independent edge models [53] and, in particular, latent space models for social networks [54, 55], we consider a generative model for social networks whose members occupy a multidimensional Blau space in section 2.1. We discuss desirable properties for social segregation statistics, and, using the generative network model, we develop a suite of statistics applicable to arbitrary attributes in section 2.2. The statistics capture segregation at different scales: single individuals, pairs of individuals, and society as a whole. Because of both their probabilistic foundations and their construction from the universal notion of a social tie, the segregation statistics have a universal scale, i.e. one unit of segregation has the same implications across different societies and at different times. We show that the segregation statistic for pairs of individuals can be a metric and can thus be used to quantify distance in Blau space. We illustrate the statistics with a simple example, and we show that it reduces to a well-known segregation statistic if the attributes are univariate and categorical: the natural logarithm of Moody's α index [34].

In section 2.4, 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 in section 3. 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 statistics, we find that physical separation and age are the most important factors contributing to the segregation of society. In section 4, we provide recommendations for conducting surveys to infer homophily in

social networks and discuss future work.

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 sex. In contrast to common latent space models [54, 55], the attributes are observed, although Hoff, Raftery and Handcock [54] 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 [56].

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 [57]. The assumption of conditionally-independent edges can be problematic. For example, it is not possible to reproduce heavy-tailed degree distributions if the node density is homogeneous and the kernel is translationally invariant [58]. Furthermore, the average degree scales linearly with the number of nodes unless the connectivity kernel ρ is adjusted to compensate [59]. Nevertheless, we use conditionally-independent edge 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 [58]. Similarly, members of the ethnic majority have more social ties in social networks in US high schools [10].

2.2. Developing a model-based segregation statistic

We have so far emphasised the desirability of metrics on social spaces with a universal characteristic scale (in the sense of being comparable between societies). After first developing universal social segregation statistics, including a notion of social separation, we will formulate both a metric and a notion of scale in section 2.3.

In addition to addressing the challenges mentioned in section 1, a social segregation statistic should satisfy the following properties: first, the statistic should be insensitive to the overall edge density to facilitate comparison of segregation across different networks. Otherwise, the segregation statistic 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 [60], we would like the statistic to capture the notion that segregation places "restrictions on the access of people to one another". Third, the statistic 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 [61] depends on the sizes of the groups even if there is no homophily: there is no natural reference point.

A single statistic cannot capture the complexities of social networks, and we develop a family of statistics applicable at different scales: (a) the *social separation* between any two individuals, (b) the *isolation* experienced by any one individual, and (c) the *social strain* experienced by society as a whole. Starting at the microscopic level, we define the *social separation* between two individuals i and j with attributes x and y as the relative log odds for j to connect with i compared to someone who is alike: the log odds intuitively capture the probabilistic nature of the conditionally independent edge model. In particular,

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

are the log odds for a connection to form with probability ρ [62]. The probability $\rho(y, y)$ for j to connect with someone who is alike serves as a reference point, and the statistic does not depend on the overall edge density. The social separation φ may be understood as the isolation experienced by i with attributes x as a result of the behaviour of j with attributes y. The statistic is zero if two individuals have the same demographic attributes or if they do not discriminate with respect to the attributes on which they differ. For a homophilous connectivity kernel, the statistic is positive and is a *semimetric* 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 homophilous, 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 *[63]: 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 statistic 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 $\rho(x, y) < \rho(y, y)$ due to homophily. □

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 [58, 23, 39, 24].

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 inferred connectivity kernels naturally propagates to the segregation statistic, as discussed in section 3. Second, we can easily

consider the properties of the segregation statistic under a variety of generative models without having to resort to computationally-expensive Monte Carlo simulations.

For example, consider a stochastic block model (SBM) [53] with K blocks, intra-group connection probability ρ_{same}, and inter-group connection probability $\rho_{\text{different}} < \rho_{\text{same}}$. Substituting into eq. (1), the social separation between two nodes with block membership x and y is

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

where 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. The social separation is equal to the natural logarithm of the α-index proposed by Moody [34] for categorical attributes x, but φ is applicable to arbitrary attributes and connectivity kernels.

The social separation $\varphi(x, y)$ is not sufficient to quantify segregation at the level of an individual: we also need to consider the distribution $P(y)$ of attributes y of all other members of society, such as their age, sex, or other demographics. For an individual with attributes x, we define the *social isolation*

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

which quantifies the average social separation between an individual with attribute x and other members of society. 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, and we have used the identity $\sum_{y=1}^{K} P(y) (1 - \delta_{xy}) = 1 - P(x)$. 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 [10]. The off-diagonal terms of the Hessian of the social isolation ϕ quantify interactions, such as the joint effect of age and ethnic differences.

To understand how segregated society is as a whole, we would like to aggregate the social isolation 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

$$\Phi = \gamma \left(\operatorname{logit} \rho_{\text{same}} - \operatorname{logit} \rho_{\text{different}} \right),$$
$$\text{where } \gamma = 1 - \sum_{x=1}^{K} P^2(x) \tag{6}$$

is an index of dispersion [34] and accounts for the relative sizes of the K blocks. The social strain is maximal when the groups are of equal size. If one of the blocks is larger, the social strain and index of dispersion approach 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 and scale in Blau space

The social separation takes a simple form if the probability for two individuals to connect is a logistic kernel [54], 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 $f_{\text{age}} = |x_{\text{age}} - y_{\text{age}}|$. Intersectionality can be accounted for by including interaction terms in the feature set. 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. According to proposition 1, the social separation $\varphi(x, y)$ is a semi-metric, and it comprises contributions from individual features, as illustrated by eq. (8). 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

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

for all l. Substituting eq. (10) into eq. (11) yields

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

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 statistic 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 [1] 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: *one unit of social separation has the same probabilistic meaning independent of the society under consideration*, facilitating comparison across disparate datasets. Even if two societies have different Blau space dimensions, e.g. a society might exist that strongly discriminates based on characteristics which are not found in other societies, the social separation between a pair of individuals has a common meaning.

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 [22], individuals they feel close to [11], or discussion partners [48, 3]. 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 [64, chapter 8]. 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 [65] 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,

$$\begin{aligned} P(f|A, \theta, I = 1) &= P(f|A, \theta) \\ &= \frac{P(A|f, \theta)P(f|\theta)}{P(A|\theta)}. \end{aligned} \tag{15}$$

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

$$\begin{aligned} P(f|\theta, I = 1) &= \sum_{\alpha=0}^{1} P(f|A = \alpha, \theta, I = 1)P(A = \alpha|\theta, I = 1) \\ &= P(f|\theta) \sum_{\alpha=0}^{1} P(A = \alpha|f, \theta) \frac{P(A = \alpha|\theta, I = 1)}{P(A = \alpha|\theta)}, \end{aligned} \tag{16}$$

where we used the identity in eq. (15) to arrive at the second line. Substituting eqs. (15) and (16) into eq. (14), the observed-data likelihood is

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

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. (17) 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 and appendix B for a validation of the inference methodology using synthetic data. For logistic connectivity kernels, the observed-data likelihood in eq. (17) resembles a conventional case-control likelihood, e.g. as used by Smith, McPherson and Smith-Lovin [17].

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 [50, 51]. 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 [22].

Demographic information about nominees can be collected either by asking seeds about their friends' demographic background [48, 3] or by conducting follow-up surveys with nominated friends [19]. The latter seems preferable because respondents may not have complete information about their social contacts. For example, the age of nominees in the British Household Panel Survey (BHPS), a dataset we consider in section 3.4, 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 C for details for each dataset). Unfortunately, follow-up surveys require additional resources to interview the nominees and may suffer from low response rates.

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 [48, 16]. In 2004, respondents were asked about the demographic background of people "with whom they discuss important matters", which tends to elicit close ties [50]. We omit all nominees who are not considered to be friends or who are family. 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 C.2. Such a complete-case analysis can introduce biases if the data are not missing completely at random, but handling the missing data in a principled fashion would require us to develop a model for demographic attributes [66].

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 C.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. For each demographic attribute, we define a feature for the logistic kernel in eq. (7), as shown in table C.1. To standardise the features $f(x_i, x_j)$, we subtract their mean and divide non-binary features by twice their standard deviation [67]; 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 [67] and simplifies the formulation of priors: we use independent, weakly-informative Cauchy priors for the kernel parameters such that

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

Following Gelman, Jakulin, Pittau and Su [68], 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 [64, 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 [69]. Summary statistics of the posterior are shown in fig. 1 (a). The connection probabilities decrease quickly with increasing age differences: the odds of connection are reduced by a multiplicative factor of about 0.3 per decade. Ethnic, sex, and religious differences all seem to have a similar effect and decrease the odds by a factor of about 0.3 each; a difference of one educational level reduces the odds by a factor of 0.7.

Hipp and Perrin [11] 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 C.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 [70]. 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" [71]. We only include nominees who are friends of seeds and exclude kinship ties; see table C.3 for details of harmonisation of attributes across seeds and nominees.

Homophily with respect to sex and ethnicity is slightly stronger than in the GSS, and educational homophily is weaker, but the inferred parameters are broadly consistent 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 [72].

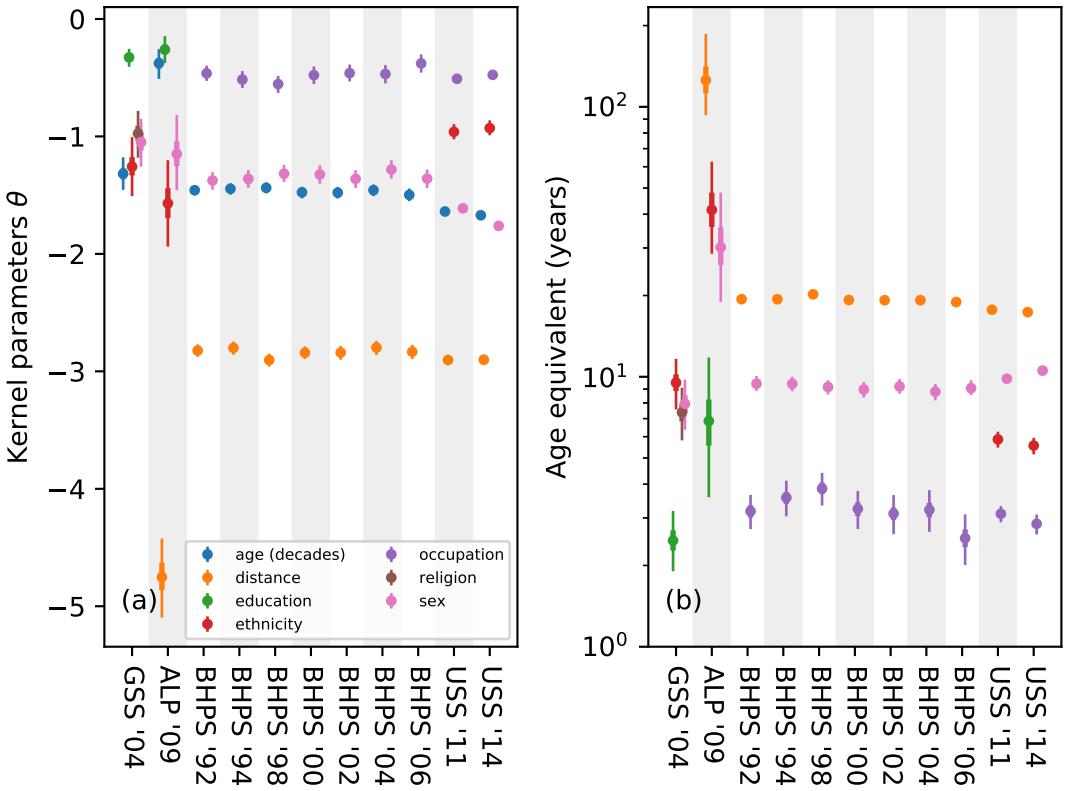

Figure 1: *Age and physical separation have a strong impact on connection probabilities, and converting parameters into age equivalents makes feature comparison more intuitive.* Panel (a) shows kernel parameters inferred from ego network data for each dataset. Panel (b) shows age equivalents. For binary 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 median, thick error bars the interquartile range, and thin error bars the 95% credible interval.

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 Survey (USS). Respondents were asked questions about "their closest friends" every other year as part of the BHPS and every three years in the USS. Data include sex, age, occupational status, ethnicity (only in the USS), and how far away friends live [73, 74] (see table C.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 [75]. Similarly, we omitted data from the BHPS in 1996 because physical separation between friends was not recorded. As shown in fig. 1, homophily seems to have increased in recent years, but the changes are likely the result of a change in methodology rather than a change in behaviour: the BHPS collected friendship information as part of the main survey, whereas the USS used a self-completion questionnaire [76].

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 GSS and USS. For each sample, we compute the social separation between pairs of respondents 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 median of the kernel parameters discussed in sections 3.2 and 3.4. We use multidimensional scaling to embed the respondents in a two-dimensional space [77], as shown in fig. 2. Panels (c) and (d) show the two-dimensional embedding that best approximates the distance matrix in the high-dimensional social space (we omit contribution due to physical space for the USS to make the embeddings comparable).

The first dimension captures the age of respondents, as illustrated in panels (a) and (b): 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 [62, chapter 6]. The second dimension captures sex and ethnicity as well as occupational status (for the USS) and education and religion (for the GSS). 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 (c) and (d) of fig. 2 also show the social isolation ϕ experienced by individuals as a greyscale heat map which we obtained in two steps: first, we evaluated an estimate

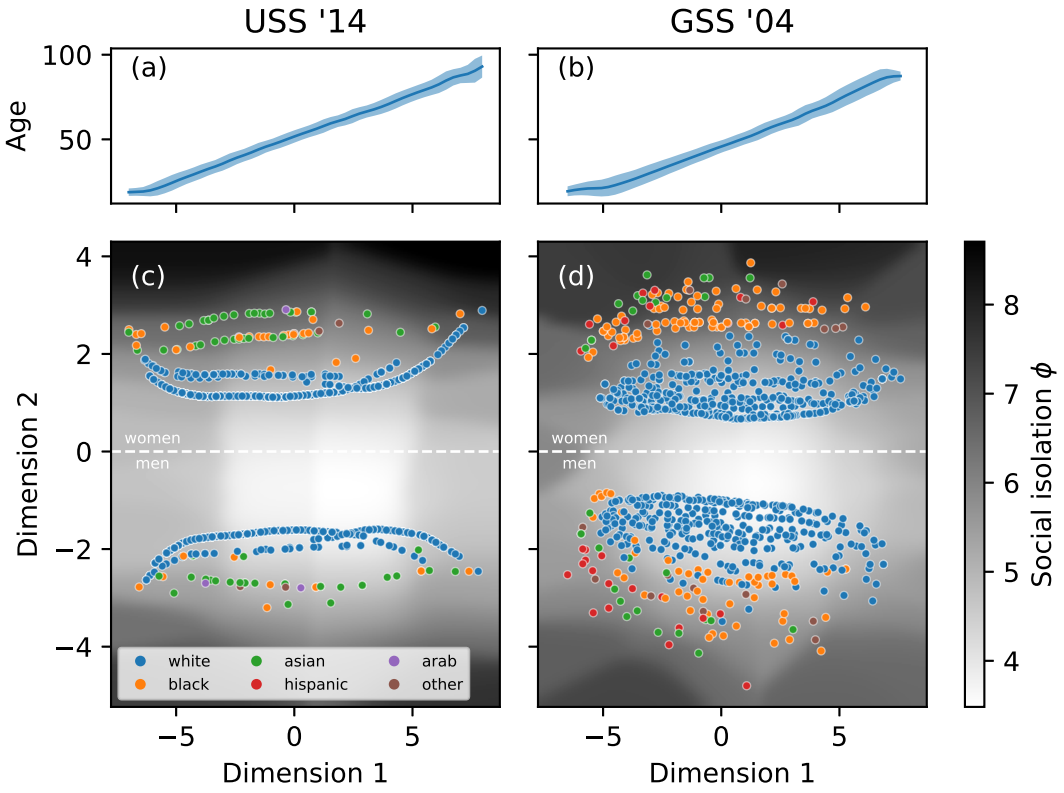

Figure 2: *A lower-dimensional embedding of the inter-node distances reveals an interpretable social space in the UK and US.* Panels (a) and (b) show the mean and standard deviation of ages as a function of the first embedding dimension as a solid line and a shaded region, respectively. Panels (c) and (d) show 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. The heat map represents a smoothed estimate of the social isolation ϕ. The "bands" of individuals in panel (c) correspond to different occupational statuses, such as employed or retired.

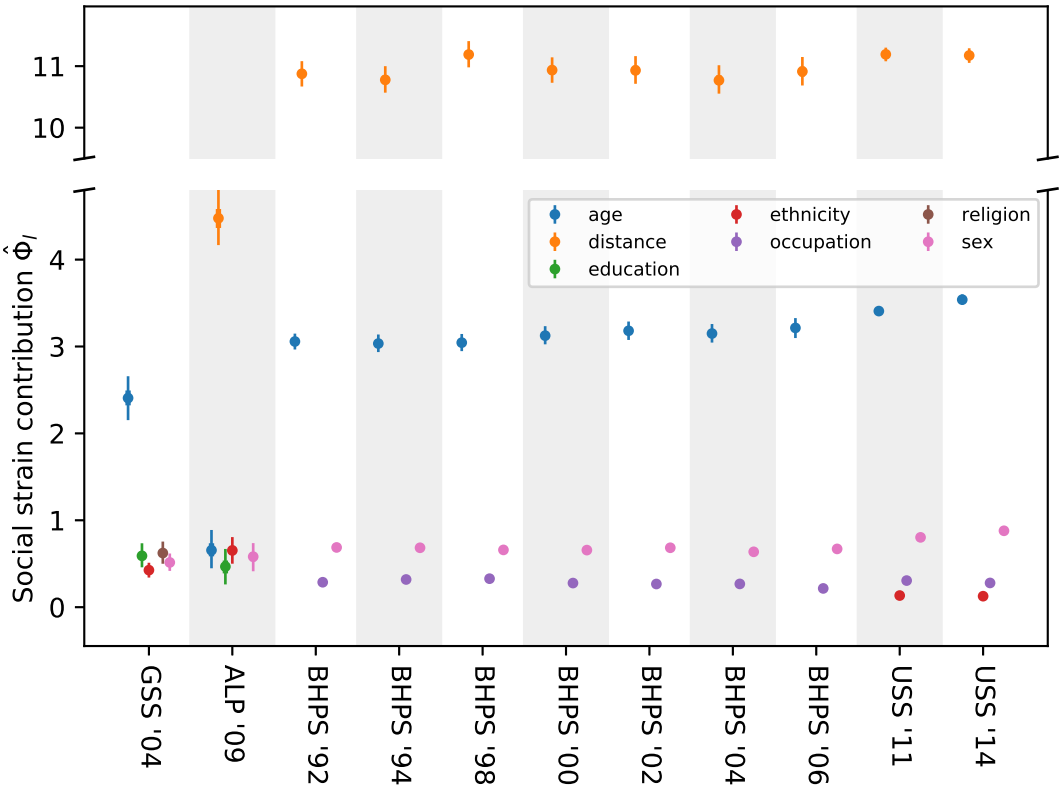

Figure 3: *Physical separation and age differences are the most important factors preventing integration of society for all datasets.* Markers represent the posterior median of the contributions to social strain for each feature, thick error bars the interquartile range, and thin error bars the 95% credible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

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. For example, members of the ethnic majority experience an average social isolation of 4.53 (4.47–4.59 95% credible interval) in the USS and 4.46 (4.10–4.84 95% credible interval) in the GSS. In contrast, the average social isolation amongst ethnic minorities is 4.99 (4.91–5.07 95% credible interval) and 5.16 (4.73–5.58 95% credible interval): significantly higher than for the ethnic majority.

Similar to the social separation in eq. (9), the social strain can be broken down into components

$$\Phi_l = \theta_l \int dx\ dy\ \left(f_l(y,y) - f_l(x,y)\right) P(x) P(y) \tag{18}$$

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

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

where the sum is over all distinct pairs of seeds U. The contribution Φ_l quantifies 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 $P(x)$ in Blau space: neither is sufficient on its own to quantify segregation. Furthermore, Φ and its contributions in eq. (18) can facilitate comparison across different datasets, as illustrated in fig. 3: they have an intuitive interpretation (the probability of edges decreases with increasing segregation) and universal scale (one unit of segregation has the same effect on edge probability).

As might be expected based on previous studies [23, 24, 25, 26, 27], physical space has by far the most significant impact on how people connect with one another. Age homophily places the second strongest restriction social connections, 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 [72] discussed in section 3.3. Age homophily is known to be particularly strong for friendship networks [2]. Homophily with respect to sex, ethnicity, education, religion, and occupation make similar, smaller contributions to the segregation of friendships. 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. We developed a family of segregation statistics with a universal scale (since they are based on the common notion of the probability of a social tie), facilitating comparison between datasets collected at different times or in different cultural contexts. Furthermore, the segregation statistics are applicable to mixed attribute types, have a natural reference point, and an intuitive interpretation: the probability to form connections decreases with increasing segregation. They are applicable at different resolutions: connections, individuals, and society as a whole. 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. The model-based approach facilitates the study of segregation in synthetic social networks, the effect of interventions, and principled quantification of uncertainties in an applied setting.

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. We used the Blau space metric to evaluate the social distance amongst respondents of the GSS and USS. 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.

The importance of physical distance highlights that our suite of segregation statistics does not distinguish between *choice homophily* and *opportunistic homophily* [78]. 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. 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. Similarly, the statistics do not discriminate between choice homophily and *social influence*, i.e. the tendency for people to become more alike given a connection [79].

Other features, including sex, ethnicity, religion, education, and occupation, have smaller effects on the presence of connections. Notably, the social strain due to ethnicity in the GSS and ALP is larger than in the USS: first, the effect of ethnicity is more pronounced in the US, as shown in fig. 1. Second, society in the US is more ethnically diverse than in the UK (79% white in GSS '04 compared with 89% white in USS '14), increasing strain on average, as exemplified with a SBM in section 2.2.

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 USS are longitudinal studies such that consistent parameter estimates are less surprising). Our observations, together with a study by Mossong et al. [4] finding that "mixing patterns [...] were remarkably similar across different European countries", suggest that connectivity kernels for friendships vary little across societies and time. 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 [51]. In particular, such surveys should explore options to explicitly incentivise nominees to provide data about themselves [80]: seeds may not recall certain attributes, or nominees may deliberately portray themselves inaccurately [81]. 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 [82]. Whenever possible, aggregation of attributes such as age into bins should be avoided because it limits the ability to infer kernel parameters [72], as we saw in section 3.3. Connectivity kernels should be inferred jointly for all dimensions of Blau space to control for social preferences on correlated attributes.

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 [10]. However, exponential random graph models may be able to better capture

the nature of social networks [83]. Furthermore, we have used a connectivity kernel that (a) is symmetric and cannot identify whether there is a status order in society [20, 56] 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 [84]—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" [85], 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 [86] because they can better capture complex patterns, such as interactions between different demographic attributes. We note that, irrespective of the choice of connectivity kernel, the interpretable segregation statistics considered here remain valid and useful.

Acknowledgements

We would like to thank Sahil Loomba and two anonymous referees for useful feedback on the manuscript and acknowledge support from the grant EP/N014529/1.

References

1. Blau PM. A Macrosociological Theory of Social Structure. Am. J. Sociol. 1977; 83:26–54. DOI: 10.1086/226505
2. McPherson M, Smith-Lovin L and Cook JM. Birds of a Feather: Homophily in Social Networks. Annu. Rev. Sociol. 2001; 27:415–44. DOI: 10.1146/annurev.soc.27.1.415
3. McPherson M, Smith-Lovin L and Brashears ME. Social Isolation in America: Changes in Core Discussion Networks over Two Decades. Am. Sociol. Rev. 2006; 71:353–75. DOI: 10.1177/000312240607100301
4. Mossong J et al. Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases. PLOS Med. 2008; 5:e74. DOI: 10.1371/journal.pmed.0050074
5. Golub B and Jackson MO. How Homophily Affects the Speed of Learning and Best-Response Dynamics. Q. J. Econ. 2012; 127:1287–338. DOI: 10.1093/qje/qjs021
6. Boutyline A and Willer R. The Social Structure of Political Echo Chambers: Variation in Ideological Homophily in Online Networks. Polit. Psychol. 2017; 38:551–69. DOI: 10.1111/pops.12337
7. Bakshy E, Messing S and Adamic LA. Exposure to ideologically diverse news and opinion on Facebook. Science 2015; 348:1130–2. DOI: 10.1126/science.aaa1160

8. DeMarzo PM, Vayanos D and Zwiebel J. Persuasion Bias, Social Influence, and Unidimensional Opinions. Q. J. Econ. 2003; 118:909–68. DOI: 10.1162/00335530360698469

9. Salathé M and Bonhoeffer S. The effect of opinion clustering on disease outbreaks. J. R. Soc. Interface 2008; 5:1505–8. DOI: 10.1098/rsif.2008.0271

10. Currarini S, Jackson MO and Pin P. An economic model of friendship: homophily, minorities, and segregation. Econometrica 2009; 77:1003–45. DOI: 10.3982/ECTA7528

11. Hipp JR and Perrin AJ. The Simultaneous Effect of Social Distance and Physical Distance on the Formation of Neighborhood Ties. City & Community 2009; 8:5–25. DOI: 10.1111/j.1540-6040.2009.01267.x

12. Wang Y, Zang H and Faloutsos M. Inferring cellular user demographic information using homophily on call graphs. *IEEE Conference (Computer Communications Workshops)*. 2013 :211–6. DOI: 10.1109/INFCOMW.2013.6562897

13. Leo Y, Fleury E, Alvarez-Hamelin JI, Sarraute C and Karsai M. Socioeconomic correlations and stratification in social-communication networks. J. R. Soc. Interface 2016; 13:20160598. DOI: 10.1098/rsif.2016.0598

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

15. Chang J, Rosen I, Backstrom L and Marlow C. ePluribus: ethnicity on social networks. *ICWSM*. 2010

16. Marsden PV. Homogeneity in confiding relations. Soc. Netw. 1988; 10:57–76. DOI: 10.1016/0378-8733(88)90010-X

17. Smith JA, McPherson M and Smith-Lovin L. Social Distance in the United States: Sex, Race, Religion, Age, and Education Homophily among Confidants, 1985 to 2004. Am. Sociol. Rev. 2014; 79:432–56. DOI: 10.1177/0003122414531776

18. Social and spatial ethnic segregation: A framework for analyzing segregation with large-scale spatial network data. *Symposium on Computing for Development*. 2013. DOI: 10.1145/2537052.2537061

19. Johnson MA. Variables Associated with Friendship in an Adult Population. J. Soc. Psychol. 1989; 129:379–90. DOI: 10.1080/00224545.1989.9712054

20. Chan TW and Goldthorpe JH. Is There a Status Order in Contemporary British Society? Evidence from the Occupational Structure of Friendship. Eur. Sociol. Rev. 2004; 20:383–401. DOI: 10.1093/esr/jch033

21. Platt L. Muslims in Britain: making social and political space. Routledge, 2012. Chap. Exploring social spaces of Muslims:53–83

22. Kalmijn M and Vermunt JK. Homogeneity of social networks by age and marital status: A multilevel analysis of ego-centered networks. Soc. Netw. 2007; 29:25–43. DOI: 10.1016/j.socnet.2005.11.008

23. Lambiotte R, Blondel VD, Kerchove C de, Huens E, Prieur C, Smoreda Z and Dooren PV. Geographical dispersal of mobile communication networks. Physica A 2008; 387:5317–25. DOI: 10.1016/j.physa.2008.05.014

24. Expert P, Evans TS, Blondel VD and Lambiotte R. Uncovering space-independent communities in spatial networks. PNAS 2011; 108:7663–8. DOI: 10.1073/pnas.1018962108

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

26. Scellato S, Noulas A, Lambiotte R and Mascolo C. Socio-spatial properties of online location-based social networks. *ICWSM*. 2011 :329–36

27. Illenberger J, Nagel K and Flötteröd G. The Role of Spatial Interaction in Social Networks. Netw. Spat. Econ. 2013; 13:255–82. DOI: 10.1007/s11067-012-9180-4

28. Rodriguez-Moral A and Vorsatz M. Complex Networks and Dynamics. Springer, 2016. Chap. An Overview of the Measurement of Segregation: Classical Approaches and Social Network Analysis:93–119. DOI: 10.1007/978-3-319-40803-3_5

29. Bojanowski M and Corten R. Measuring segregation in social networks. Soc. Netw. 2014; 39:14–32. DOI: 10.1016/j.socnet.2014.04.001

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

31. Popielarz PA. (In)voluntary Association: A Multilevel Analysis of Gender Segregation in Voluntary Organizations. Gender & Society 1999; 13:234–50. DOI: 10.1177/089124399013002005

32. Charles M and Grusky DB. Models for Describing the Underlying Structure of Sex Segregation. Am. J. Sociol. 1995; 100:931–71. DOI: 10.1086/230605

33. Reardon SF and O'Sullivan D. Measures of Spatial Segregation. Sociological Methodology 2004; 34:121–62. DOI: 10.1111/j.0081-1750.2004.00150.x

34. Moody J. Race, School Integration, and Friendship Segregation in America. Am. J. Sociol. 2001; 107:679–716. DOI: 10.1086/338954

35. Newman MEJ. Mixing patterns in networks. Phys. Rev. E 2003; 67:026126. DOI: 10.1103/PhysRevE.67.026126

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

37. Kim M and Leskovec J. Multiplicative Attribute Graph Model of Real-World Networks. Internet Mathematics 2012; 8:113–60. DOI: 10.1080/15427951.2012.625257

38. Pelechrinis K and Wei D. VA-Index: Quantifying Assortativity Patterns in Networks with Multidimensional Nodal Attributes. PLOS ONE 2016; 11:e0146188. DOI: 10.1371/journal.pone.0146188

39. Butts CT, Acton RM, Hipp JR and Nagle NN. Geographical variability and network structure. Soc. Netw. 2012; 34:82–100. DOI: 10.1016/j.socnet.2011.08.003

40. Golder SA and Macy MW. Digital Footprints: Opportunities and Challenges for Online Social Research. Annu. Rev. Sociol. 2014; 40:129–52. DOI: 10.1146/annurev-soc-071913-043145

41. Blumenstock J, Cadamuro G and On R. Predicting poverty and wealth from mobile phone metadata. Science 2015; 350:1073–6. DOI: 10.1126/science.aac4420

42. Luo S, Morone F, Sarraute C, Travizano M and Makse HA. Inferring personal economic status from social network location. 2017; 8:15227. DOI: 10.1038/ncomms15227

43. Wang Y and Kosinski M. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. Open Science Framework 2017 :zn79k

44. Kosinski M, Stillwell D and Graepel T. Private traits and attributes are predictable from digital records of human behavior. PNAS 2013; 110:5802–5. DOI: 10.1073/pnas.1218772110

45. Backstrom L, Dwork C and Kleinberg J. Wherefore Art Thou R3579X?: Anonymized Social Networks, Hidden Patterns, and Structural Steganography. Communications of the ACM 2011; 54:133–41. DOI: 10.1145/2043174.2043199

46. Narayanan A and Shmatikov V. Robust de-anonymization of Large Sparse Datasets. *Symposium on Security and Privacy*. 2008 :111–25. DOI: 10.1109/SP.2008.33

47. Huckfeldt RR. Social Contexts, Social Networks, and Urban Neighborhoods: Environmental Constraints on Friendship Choice. Am. J. Sociol. 1983; 89:651–69. DOI: 10.1086/227908

48. Marsden PV. Core Discussion Networks of Americans. Am. Sociol. Rev. 1987; 52:122–31. DOI: 10.2307/2095397

49. Banerjee A, Chandrasekhar AG, Duflo E and Jackson MO. The Diffusion of Microfinance. Science 2013; 341:1236498. DOI: 10.1126/science.1236498

50. Marin A. Are respondents more likely to list alters with certain characteristics? Implications for name generator data. Soc. Netw. 2004; 26:289–307. DOI: 10.1016/j.socnet.2004.06.001

51. Eagle DE and Proeschold-Bell RJ. Methodological considerations in the use of name generators and interpreters. Soc. Netw. 2015; 40:75–83. DOI: 10.1016/j.socnet.2014.07.005

52. Eveland Jr. WP, Appiah O and Beck PA. Americans are more exposed to difference than we think: Capturing hidden exposure to political and racial difference. Soc. Netw. 2017; 52:192–200. DOI: 10.1016/j.socnet.2017.08.002

53. Snijders TA. Statistical Models for Social Networks. Annu. Rev. Sociol. 2011; 37:131–53. DOI: 10.1146/annurev.soc.012809.102709

54. Hoff PD, Raftery AE and Handcock MS. Latent Space Approaches to Social Network Analysis. J. Am. Stat. Assoc. 2002; 97:1090–8. DOI: 10.1198/016214502388618906

55. Hoff PD. Multiplicative latent factor models for description and prediction of social networks. Comput. Math. Organ. Theory 2008; 15:261–72. DOI: 10.1007/s10588-008-9040-4

56. Ball B and Newman M. Friendship networks and social status. Network Science 2013; 1:16–30. DOI: 10.1017/nws.2012.4

57. Fienberg SE. A Brief History of Statistical Models for Network Analysis and Open Challenges. J. Comput. Graph. Stat. 2012; 21:825–39. DOI: 10.1080/10618600.2012.738106

58. Barnett L, Di Paolo E and Bullock S. Spatially embedded random networks. Phys. Rev. E 2007; 76:056115. DOI: 10.1103/PhysRevE.76.056115

59. Caron F and Fox EB. Sparse graphs using exchangeable random measures. J. R. Stat. Soc. B 2017; 79:1295–366. DOI: 10.1111/rssb.12233

60. Freeman LC. Segregation in Social Networks. Sociol. Methods Res. 1978; 6:411–29. DOI: 10.1177/004912417800600401

61. Krackhardt D and Stern RN. Informal Networks and Organizational Crises: An Experimental Simulation. Soci. Psychol. Q. 1988; 51:123–40. DOI: 10.2307/2786835

62. Hastie T, Tibshirani R and Friedman J. The elements of statistical learning: data mining, inference and prediction. Springer, 2009. DOI: 10.1007/978-0-387-84858-7

63. Wilson WA. On Semi-Metric Spaces. Am. J. Math. 1931; 53:361–73. DOI: 10.2307/2370790

64. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A and Rubin DB. Bayesian Data Analysis. Chapman and Hall/CRC, 2013

65. King G and Zeng L. Logistic Regression in Rare Events Data. Political Anal. 2001; 9:137–63. DOI: 10.1093/oxfordjournals.pan.a004868

66. Pigott TD. A Review of Methods for Missing Data. Educ. Res. Eval. 2001; 7:353–83. DOI: 10.1076/edre.7.4.353.8937

67. Gelman A. Scaling regression inputs by dividing by two standard deviations. Stat. Med. 2008; 27:2865–73. DOI: 10.1002/sim.3107

68. Gelman A, Jakulin A, Pittau MG and Su YS. A weakly informative default prior distribution for logistic and other regression models. Ann. Appl. Stat. 2008; 2:1360–83. DOI: 10.1214/08-AOAS191

69. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 1970 :97–109. DOI: 10.1093/biomet/57.1.97

70. Pollard M and Baird MD. The RAND American Life Panel. Tech. rep. RAND, 2017

71. Mihaly K. American Life Panel: Well-being 86 questionnaire. Tech. rep. RAND corporation, 2009

72. Hutcheon JA, Chiolero A and Hanley JA. Random measurement error and regression dilution bias. BMJ 2010; 340:1402–6. DOI: 10.1136/bmj.c2289

73. Institute for Social and Economic Research. British Household Panel Survey Questionnaire Wave 10. 2000. Available from: https://www.iser.essex.ac.uk/bhps/documentation/pdf_versions/questionnaires/bhpsw10q.pdf

74. Institute for Social and Economic Research. UK Household Longitudinal Study Mainstage Questionnaire Wave 3. 2017. Available from: https://www.understandingsociety.ac.uk/documentation/mainstage/questionnaire/questionnaire-documents/mainstage/wave-3/Understanding_Society_Wave_3_Questionnaire_v03.pdf

75. Unusual age distribution after conditioning on wJBSTATT in wave R. 2016. Available from: https://www.understandingsociety.ac.uk/support/issues/687

76. Unexpectedly strong gender homophily in Understanding Society compared with the BHPS. 2017. Available from: https://www.understandingsociety.ac.uk/support/issues/869

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

78. Franz S, Marsili M and Pin P. Observed Choices And Underlying Opportunities. Science and Culture 2010; 76:471–6

79. Shalizi CR and Thomas AC. Homophily and Contagion Are Generically Confounded in Observational Social Network Studies. Sociol. Methods Res. 2011; 40:211–39. DOI: 10.1177/0049124111404820

80. Biernacki P and Waldorf D. Snowball Sampling: Problems and Techniques of Chain Referral Sampling. Sociol. Methods Res. 1981; 10:141–63. DOI: 10.1177/004912418101000205

81. Bruch E, Feinberg F and Lee KY. Extracting multistage screening rules from online dating activity data. PNAS 2016; 113:10530–5. DOI: 10.1073/pnas.1522494113

82. Po JYT, Finlay JE, Brewster MB and Canning D. Estimating Household Permanent Income from Ownership of Physical Assets. Tech. rep. 97. Harvard, 2012

83. Wimmer A and Lewis K. Beyond and Below Racial Homophily: ERG Models of a Friendship Network Documented on Facebook. Am. J. Sociol. 2010; 116:583–642. DOI: 10.1086/653658

84. Bhattacharya K, Ghosh A, Monsivais D, Dunbar RIM and Kaski K. Sex differences in social focus across the life cycle in humans. Open Science 2016; 3:160097. DOI: 10.1098/rsos.160097

85. Stouffer SA. Intervening Opportunities: A Theory Relating Mobility and Distance. Am. Sociol. Rev. 1940; 5:845–67. DOI: 10.2307/2084520

86. Frölich M. Non-parametric regression for binary dependent variables. Econom. J. 2006; 9:511–40. DOI: 10.1111/j.1368-423X.2006.00196.x

The code to reproduce the results and figures is available at https://github.com/tillahoffmann/kernels.

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 [1]. Including weights in Bayesian analyses is generally difficult [2], and, in principle, we should model the data collection process explicitly [3, chapter 8]. Unfortunately, modelling the data collection process is non-trivial, and we use a weighted pseudo-likelihood instead [4]. In particular, the observed-data log-likelihood from eq. (17) 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 [1].

A.2. Numerical stability

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

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

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

is the logistic function. In particular, note that $1 - \sigma(\xi) = \sigma(-\xi)$ and $\log \sigma(\xi) = -\operatorname{log1p}\left[\exp(-\xi)\right]$, where $\operatorname{log1p}(\xi) = \log(1 + \xi)$ is a numerically stable implementation—even for $|\xi| \ll 1$. Substituting into eq. (17) yields

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

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

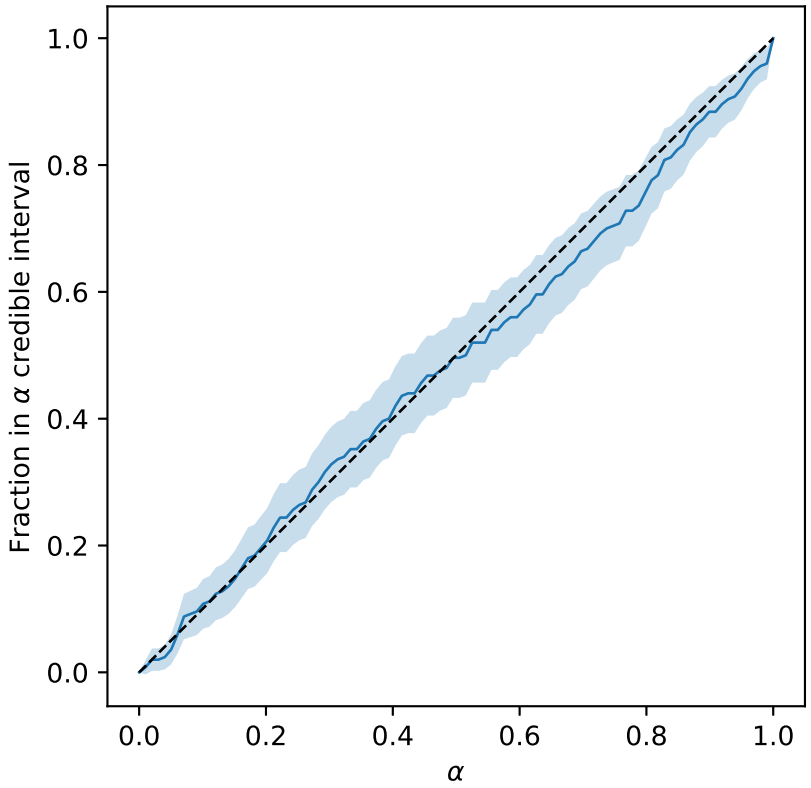

Figure B.1: *A coverage analysis of posterior credible intervals validates the inference methodology.* The blue line shows the fraction of inferences for which the true parameter values are contained in the α-credible interval of a Laplace approximation of the posterior. The shaded region corresponds to two standard deviations of the mean across 250 simulations.

B. Validation of inference methodology using synthetic ego network data

To test the inference methodology, we conduct a coverage analysis of posterior credible intervals in three steps: fist, we generate 250 synthetic ego network datasets with known kernel parameter values. Second, we infer the parameter posterior distribution. Third, we evaluate the proportion $\lambda(\alpha)$ of true parameter values contained in the α-credible interval across multiple synthetic datasets. We expect the true parameter values to be contained in the α-credible interval for a proportion α of the synthetic datasets [3, section 10.7], i.e. $\lambda(\alpha) \approx \alpha$.

In the first step, we draw the positions x of $n = 2,000$ nodes uniformly at random from the unit square, and we connect nodes to one another according to a logistic connectivity

kernel with features

$$f(x_i, x_j) = \left(1, \frac{3\left[|x_{i1} - x_{j1}| - \frac{1}{3}\right]}{\sqrt{2}}, \frac{3\left[|x_{i2} - x_{j2}| - \frac{1}{3}\right]}{\sqrt{2}}\right),$$

where the first feature represents the bias, and the last two features capture distance in Blau space. The features were chosen to be standardised in the same fashion as described in section 3.2, i.e. to have zero mean and a standard deviation of 0.5 [5]. The corresponding parameter values θ are drawn from a normal distribution with unit variance and expectation

$$\langle\theta\rangle = (-7, 0, 0)\,.$$

The expectation $\langle\theta\rangle$ was chosen such that it gives rise to a typical degree of $2,000 \times \sigma(-7) \approx 1.8$, similar to the real-world datasets considered in section 3. We select $s = 100$ respondents as egos and include all their alters as positive examples. We sample three times as many negative examples by selecting distinct pairs of respondents. If the number of respondents is not sufficient to draw the desired number of distinct control pairs, we use all $\frac{s(s-1)}{2}$ possible distinct pairs of respondents as negative examples.

In the second step, we maximise the log-posterior using a gradient ascent algorithm to obtain the MAP estimate θ and consider the Laplace approximation of the posterior [6, section 4.4], i.e. a multivariate normal approximation in the vicinity of the MAP estimate. We evaluate the Hessian H of the negative log-posterior at the MAP estimate to obtain the precision matrix of the Laplace approximation. We do not draw samples from the posterior because the Laplace approximation is computationally more convenient.

In the last step, we consider the quantity

$$\chi^2 = \left(\theta - \hat{\theta}\right)^\intercal H \left(\theta - \hat{\theta}\right) \tag{B.1}$$

which we expect to follow a χ^2 distribution with three degrees of freedom [7]. We calculate the statistic in eq. (B.1) for each simulation and consider the empirical probability that χ^2 does not exceed the expected quantiles of the χ^2-distribution, as shown in fig. B.1. As expected, the α-credible interval contains the true parameter values for a fraction α of inferences, validating the inference methodology.

C. Coding of demographic attributes and feature maps

In the BHPS and USS, 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

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

census reporting areas and have a few thousand inhabitants each [8]. 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. The USS furthermore distinguishes between friends living more than fifty miles apart but within the UK and friends outside the UK. We discard the latter (2.6% and 2.1% of all friends in waves C and F of the USS) because it is difficult to define an appropriate control population. For the BHPS, we implicitly assume that all friends are resident in the UK.

In the USS, 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 quantify 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, $f_{\text{ethnicity}}(x_i, x_j) = 0.5$.

References

1. Kish L. Weighting for unequal P_i. J. Off. Stat. 1992; 8:183–200
2. Gelman A. Struggles with survey weighting and regression modeling. Statistical Science 2007; 22:153–64. DOI: 10.1214/088342306000000691
3. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A and Rubin DB. Bayesian Data Analysis. Chapman and Hall/CRC, 2013
4. Pfeffermann D. The use of sampling weights for survey data analysis. Stat. Methods Med. Res. 1996; 5:239–61. DOI: 10.1177/096228029600500303
5. Gelman A. Scaling regression inputs by dividing by two standard deviations. Stat. Med. 2008; 27:2865–73. DOI: 10.1002/sim.3107
6. Bishop C. Pattern Recognition and Machine Learning. Springer, 2006
7. Slotani M. Tolerance regions for a multivariate normal population. Ann. Inst. Stat. Math. 1964; 16:135–53. DOI: 10.1007/BF02868568

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 C.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	Egos	Dropped egos	Alters	Dropped alters
GSS '04	2,774	38 (1.4%)	863	158 (15.5%)
ALP '09	2,472	0 (0.0%)	2,481	315 (11.3%)
BHPS '92	9,105	1 (<0.1%)	18,219	506 (2.7%)
BHPS '94	8,728	5 (0.1%)	17,328	469 (2.6%)
BHPS '98	8,584	5 (0.1%)	16,949	565 (3.2%)
BHPS '00	8,281	2 (<0.1%)	16,255	432 (2.6%)
BHPS '02	7,971	0 (0.0%)	15,716	502 (3.1%)
BHPS '04	7,609	0 (0.0%)	14,971	502 (3.2%)
BHPS '06	7,459	0 (0.0%)	14,558	331 (2.2%)
USS '11	36,526	199 (0.5%)	74,141	461 (0.6%)
USS '14	29,082	319 (1.1%)	61,892	611 (1.0%)

Table C.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.* Individuals excluded for other reasons, e.g. due to being a relative or under the age of 18, are not listed.

8. Department for Communities and Local Government. Lower layer super output areas. Available from: http://opendatacommunities.org/data/lower-layer-super-output-areas

9. Smith JA. A Social Space Approach to Testing Complex Hypotheses: The Case of Hispanic Marriage Patterns in the United States. Socius 2017; 3:2378023117739176. DOI: 10.1177/2378023117739176

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 C.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 C.3. The joint effect of ethnic differences and whether people identify as Hispanic is still unclear [9]; 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 reported in the survey as a regression feature and generate control features using Monte Carlo simulation, as discussed in appendix C.
[c]See appendix C for a detailed description of the feature map.

Table C.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.