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 structure of social networks can have a profound impact on our daily lives: social capital, segregation, and processes on networks, such as opinion formation and the spread of pathogens, are all affected by the web of interactions between individuals who tend to associate with others who are alike: a phenomenon known as homophily.

We fuse the notion of Blau space—a space spanned by the attributes of the members of society that inhabit it—with conditionally-independent edge models to obtain a generative model for social networks. Under a Bayesian paradigm, we develop the techniques to infer the parameters of the model from ego network data. The model-based approach allows us to develop a social segregation measure with desirable properties, including an intuitive scale, being applicable to multivariate and mixed node attributes, as well as being defined at the level of individuals and the society as a whole. For pairs of individuals, the segregation measure induces a semi-metric that can be used to measure the social distance between them in a principled fashion.

We apply our approach to ego network data collected in the United Kingdom and United States from four surveys, and make recommendations for future ego-network surveys. 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 social distance 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 [**McPherson2001**]. Whilst institutionalised segregation and discrimination have

largely been outlawed in the developed world, voluntary association between individuals who share common traits makes social circles more homogeneous than one would expect by chance. 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 on the social fabric [**Golub2012**]. For example, users of online social networks such as Facebook and Twitter tend to connect with others who hold similar political views [**Boutyline2017**]. They are more likely to be exposed to information that confirms rather than challenges their beliefs [**Bakshy2015**]. An "echo chamber" effect ensues, leading to polarised opinions [**DeMarzo2003**]. Homophily can also have a detrimental impact on public health: **Salathe2008** showed that clusters of individuals who mutually reinforce their belief that vaccinations are harmful can raise the likelihood of significant disease outbreaks—even if the vaccination rate is above herd immunity levels on average.

Homophily can be observed in friendships [**Currarini2009**, **Hipp2009**], networks of discussion partners [**McPherson2006**], communication networks [**Wang2013**, **Leo2016**], marital ties [**Blau1984**], and online social networks [**Chang2010**]. Relationships are homogeneous with respect to a wide range of attributes such as age [**Marsden1988**, **Smith2014**], gender [**Stehle2011**, **Smith2014**], ethnicity [**Chang2010**, **Blumenstock2013**, **Currarini2009**], education [**McPherson2006**, **Smith2014**, **Johnson1989**], occupation [**Chan2004**], income [**Leo2016**, **Wang2013**, **Johnson1989**], religion [**Platt2012**], parental [**Johnson1989**] and marital status [**Kalmijn2007**], political ideology [**Bakshy2015**, **Boutyline2017**], and geographical location [**Lambiotte2008**, **Expert2011**, **Backstrom2010**, **Scellato2011**, **Illenberger2013**]. Romantic relationships, which tend to be heterophilous with respect to gender, are a notable exception.

The study of homophily is complicated by the scarcity of high-quality data [**Butts2012**, **Blumenstock2013**]: we need social network data together with demographic information for each person. On the one hand, online social networks and the widespread use of mobile phones provide us with detailed information about connections between individuals [**Golder2014**], and seemingly private traits such as socioeconomic status [**Blumenstock2015**, **Luo2017**], sexual orientation [**Wang2017**], age, gender, and political ideology can be inferred [**Kosinski2013**]. Unfortunately, network features are often used to predict demographic attributes [**Wang2013**, **Blumenstock2015**, **Luo2017**, **Kosinski2013**], which would confound any study of homophily: we cannot use the same data to predict demographic attributes and study homophily with respect to those attributes. Furthermore, data are too revealing in terms of privacy but, at the same time, do not provide enough information for researchers [**Golder2014**]. Individuals can be identified in anonymised social networks [**Backstrom2011**, **Narayanan2008**], and augmenting the network data with demographic information would make re-identification even easier.

On the other hand, 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 [**Huckfeldt1983**, **Johnson1989**], discussion partners [**Marsden1987**, **McPherson2006**], or support networks [**Kalmijn2007**, **Banerjee2013**]. The questions used to elicit social ties provide

an imperfect observation of the immediate neighbourhood of respondents [**Marin2004**, **Eagle2015**, **Eveland-Jr.2017**].

Building on the successes of latent space models for social networks [**Hoff2002**, **Handcock2007**, **Raftery2012**], we consider a generative model for social networks whose members occupy a multidimensional Blau space. We derive the posterior for model parameters given partial observations of such social networks obtained from surveys and show that the posterior is equivalent to the widely-known "prior correction" when the probability for two people to connect is modelled by a logistic model. We apply our approach to nine existing datasets from the United Kingdom (UK) and two from the United States (US). 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. 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 [**Pfeiffer-III2014**, **Lieberman2010**, **Lieberman2012**, **Nettleton2016**, **Sathanur2017**].

We also discuss desirable properties for social segregation measures and, using the generative network model, develop a suite of measures applicable to arbitrary attributes. The measures capture segregation at different scales: pairs of individuals, single 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. We apply the segregation measures to survey data, finding that physical separation and age are the most important factors contributing to the segregation of society.

2 Methods

2.1 Generative network model and observation

We build a generative model for ego networks in two steps: we discuss a model for entire social networks first and the data collection process second. In particular, we consider 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 [**Hoff2002**, **Hoff2008**, **Handcock2007**], the attributes are observed although **Hoff2002** also consider observed covariates. This allows us to learn how members of the population connect with one another. In particular, the q-dimensional attribute vector $x_i \in \mathbb{B}$ for each individual $i \in N$ is drawn independently from a distribution of demographic attributes $P(x_i)$. 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 directed because social ties need not be reciprocated. For example, Alice may consider Bob to be a friend whereas the reverse need not be true [**Ball2013**].

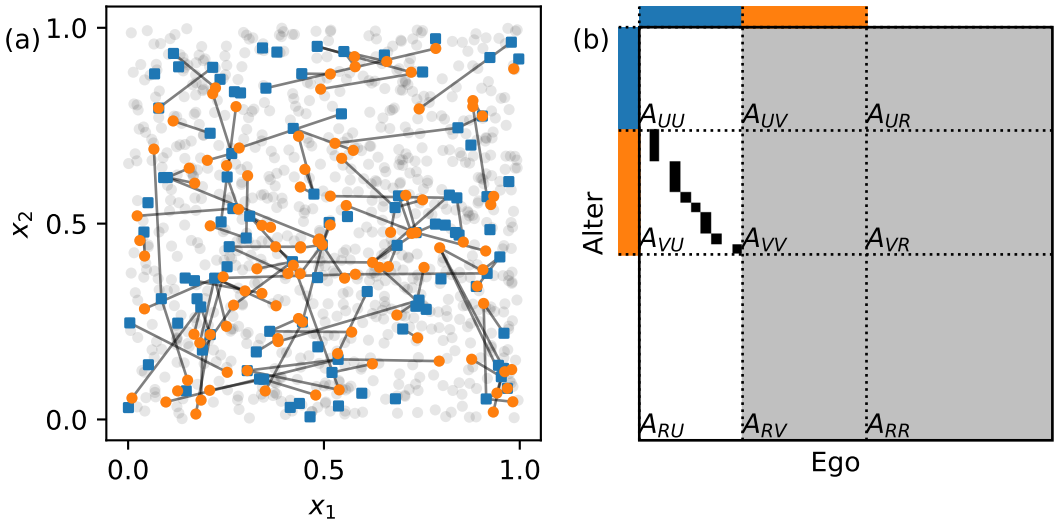

Figure 1: *A simple generative model for ego networks obtained from surveys.* Panel (a) shows synthetic survey data generated by our model. Seeds U are represented by blue squares, nominees V by orange circles, and the remainder of the population R by grey circles. Panel (b) illustrates the parts of the adjacency matrix revealed by a particular two-stage survey. Connections are represented by black entries in the adjacency matrix, absent connections by white entries, and lacking information by grey entries.

Given the positions of two individuals i and j in Blau space, we assume that connections form independently with probability $\rho(x_i, x_j, \theta)$, i.e. we consider a spatial network model [**Barnett2007**] with conditionally-independent edges [**Fienberg2012**]. The *connectivity kernel* ρ is parametrised by a vector of parameters θ which we would like to infer. The assumption of conditionally-independent edges can be problematic. For example, it is difficult to reproduce some of the statistics of real-world networks, such as heavy-tailed degree distributions, if the node density is homogeneous [**Barnett2007**]. Furthermore, keeping θ constant, the average degree scales linearly with the number of nodes [**Caron2017**]. Nevertheless, we use spatial network models because the connectivity kernel is easily interpretable and they are able to capture salient features of social networks. For example, nodes in high-density regions have larger degrees on average [**Barnett2007**]. Similarly, members of the ethnic majority have more social ties in social networks in US high schools [**Currarini2009**].

We now turn to the data collection process. Observations are collected in two stages: first, we sample a subset $U \subset N$ of *seeds* uniformly at random and collect information about their demographic background. We also ask each seed $j \in U$ about their social ties $V_j = \{i \in N : A_{ij} = 1\}$, henceforth *nominees*. Second, we collect demographic information about the nominees $V = \bigcup_{j \in U} V_j$ by asking the seeds about their friends [**Huckfeldt1983**, **McPherson2006**, **Kalmijn2007**] or conducting follow-up surveys [**Johnson1989**].

Panel (a) of **??** illustrates the model with a simple example: we first draw the positions x of $n = 1,000$ nodes in the social space uniformly at random from the two-dimensional unit box. Second, we connect nodes with each another using a connectivity kernel that induces homophily along the two dimensions of Blau space. Finally, we sample $|U| = 100$ seeds from the population, and collect information about their demographic background as well as their $|V| = 108$ nominees, which allows us to embed the ego networks in Blau space.

2.2 Inference

[TAH: big gaping hole here]

2.3 Developing a segregation measure

The study of homophily is intimately tied to measuring segregation, and a wide range of measures have been developed (see **Rodriguez-Moral2016**, **Bojanowski2014** for reviews). Many approaches are based on co-presence in organisational units such as schools [**Orfield2014**], voluntary associations [**Popielarz1999**], occupations [**Charles1995**], or census tracts [**Reardon2004**, **Reardon2011**], 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 to any data that can be stratified according to a variable of interest, they cannot capture segregation at smaller scales than the strata [**Blumenstock2013**]. 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* the schools may exhibit strong ethnic homophily [**Currarini2009**, **Moody2001**]. Organisational measures cannot capture such social segregation.

Social measures of segregation, such as the assortativity coefficient [**Newman2003a**], overcome these limitations by explicitly considering the interactions amongst individuals [**Blumenstock2013**], but have their own difficulties: first, they usually rely on the existence of mutually exclusive groups such as sex, ethnicity, or religion [**Bojanowski2014**], and they are not applicable to continuous attributes such as age or income. Attributes are often discretised [**Lam-Morgan2012**, **Kalmijn2007**, **Kim2012**], but the boundaries between categories always suffer from some degree of arbitrariness [**Reardon2004**]. 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, and it is non-trivial to apply them to sampled network data.

In addition to addressing the challenges above, a desirable 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 **Freeman1978**, 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 **Krackhardt1988** 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 will develop a suite of measures applicable at different scales: the segregation or separation between any two individuals, the isolation experienced by any one individual, and a global measure of segregation. Starting at the microscopic level, we use the generative model discussed in **??** to define a *pairwise* measure of separation between two individuals with attributes x and y

$$\varphi(x, y) = -\log \frac{\rho(x, y)}{\rho(y, y)}, \tag{1}$$

where $\rho(x, y)$ is the probability for y to connect with x, and, for notational convenience, we use x to refer to an individual with attributes x as well as the attributes themselves. The probability $\rho(y, y)$ for y to connect with someone with identical attributes serves as a reference point such that the measure does not depend on the overall edge density. An individual y is $\exp(\varphi(x, y))$ times less likely to connect with someone with attributes x than they are to connect with someone with identical attributes y, and φ may be understood as the social isolation experienced by x as a result of the behaviour of y.

The measure is zero if people do not discriminate with respect to the attributes or if two individuals have the same demographic attributes. In a heterophilous society, the measure is negative for any pair of individuals. 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 **??**.

Proposition 1. *If the connectivity kernel is homophilous, symmetric, and homogeneous, the pairwise segregation measure is a* semi-metric *[**Wilson1931**], i.e. it satisfies the properties of a metric except the triangle inequality.*

Proof. First, the measure is zero for any two individuals with the same attributes by substitution into **??**. Second, the measure is non-negative because homophily implies that $\rho(x, y) < \rho(y, y)$. Third, the numerator of **??** is invariant under exchange of the arguments, and the denominator is constant because the kernel is symmetric and homogeneous, respectively. The triangle inequality is not necessarily satisfied. □

The less likely two people are to connect, the larger the social distance between them. The assumptions required for **??** to hold may seem restrictive, but they are satisfied by most studies of spatial networks [**Butts2012**, **Lambiotte2008**, **Expert2011**].

Defining the measure in terms of the 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) [**Snijders2011**] with intra-group connection probability ρ_{same} and inter-group connection probability $\rho_{\text{different}} < \rho_{\text{same}}$. The pairwise segregation measure

$$\varphi(x, y) = -(1 - \delta_{xy}) \log \frac{\rho_{\text{different}}}{\rho_{\text{same}}},$$

where x and y denote the blocks and δ_{xy} is the Kronecker delta. The segregation measure only depends block membership, and it does not depend on the size of each block. The pairwise measure for members of different block is the log odds ratio for the existence of intra-group ties as opposed to inter-group ties, which has been considered previously by **Moody2001**.

Both the connectivity kernel and the pairwise measure convey information about how two individuals connect with one another. To quantify segregation at the level of an individual x, we define the *individual* segregation measure or *social isolation*

$$\phi(x) = \int dy\ P(y)\varphi(x, y), \tag{2}$$

which measures the average distance from members of society to x. For our SBM,

$$\phi(x) = -(1 - P(x)) \log \frac{\rho_{\text{different}}}{\rho_{\text{same}}}, \tag{3}$$

and 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 [**Currarini2009**].

To understand how segregated society is as a whole, we would like to aggregate the individual measure 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 $\max_x \phi(x)$. We take a utilitarian approach and, in line with **??**, define the *global* segregation measure as

$$\Phi = \int dx\ P(x)\phi(x), \tag{4}$$

which quantifies the average social distance amongst members of the society and can be viewed as an indicator of social strain in society. For the SBM,

$$\Phi = -\gamma \log \frac{\rho_{\text{different}}}{\rho_{\text{same}}},$$
$$\text{where } \gamma = \sum_{x=1}^{K} P(x)\left(1 - P(x)\right) \tag{5}$$

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

2.4 Distance in Blau space

The segregation measure takes a simple form if the connectivity kernel is logistic and the probability of connection is small: we can approximate the logistic sigmoid such that

$$\rho(x, y, \theta) = \sigma\left(\theta^{\intercal} f(x, y)\right) \approx \exp\left(\theta^{\intercal} f(x, y)\right).$$

This approximation is virtually always applicable because the fractional approximation error is at most $\rho(x, y, \theta)$, and $\rho(x, y, \theta) \ll 1$ for social networks with a large number of nodes. The social distance between x and y is thus made up of contributions from different features of the logistic kernel, i.e.

$$\varphi(x, y) \approx \sum_{l=1}^{p} \varphi_l(x, y), \tag{6}$$
$$\text{where } \varphi_l(x, y) = -\theta_l\left(f_l(x, y) - f_l(y, y)\right) \tag{7}$$

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

Proposition 2. *The pairwise measure is a metric if the kernel is homophilous, i.e.* $\theta_l < 0$*, the edge density is small such that the approximation in* **??** *holds, and each feature* $f_l(x, y)$ *is an affine transform of a metric* $d_l(x, y)$ *for* $l > 1$*, i.e.*

$$f_l(x, y) = a_l d_l(x, y) + b_l, \tag{8}$$

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

Proof. The pairwise measure is a semi-metric according to **??**, and, according to **??**, it consists of contributions due to 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

$$\varphi_l(x, z) \geq \varphi_l(x, y) + \varphi_l(y, z). \tag{9}$$

Substituting **??** into **??** yields

$$-\theta_l a_l d_l(x, z) \geq -\theta_l a_l \left[d_l(x, y) + d_l(y, z)\right], \tag{10}$$

where we have used the metric property $d_l(x, x) = 0$ for all x and the constant b_l in **??** vanishes by **??**. The inequality in **??** holds because $\theta_l < 0$, $a_l > 0$ by assumption, and $d_l(x, y)$ is a metric. □

In other words, the pairwise measure is a true measure of distance in the social space with a probabilistic interpretation if the edge density is low and the features are themselves measures of distance. This observation puts Peter Blau's [**Blau1977**] 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 theoretical footing: fitting conditionally-independent edge models allows us to learn the metric of Blau space.

3 Application

3.1 Ego network data collected in surveys

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 [**Kalmijn2007**], individuals they feel close to [**Hipp2009**], or discussion partners [**Marsden1987**, **McPherson2006**]. The recovered social ties depend on the nature of the relationship, the mode of administration of the questionnaire (e.g. face-to-face, telephone interview, or online survey) as well as the interviewer [**Marin2004**, **Eagle2015**]. Consequently, we do not expect the kernel parameters inferred from different datasets to be 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 *voluntary* association amongst members of the population rather than the social structures they were born into [**Kalmijn2007**].

Demographic information about nominees can be collected either by asking seeds about their friends' demographic background [**Marsden1987**, **McPherson2006**] or by conducting follow-up surveys [**Johnson1989**]. 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. 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. Because the researchers conducting the surveys anticipated such difficulties, 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, a process we discuss in the context of each individual dataset.

We need to address two additional challenges before we can apply the inference algorithm to real data: first, individuals are often not included in the survey uniformly at random, and weights are traditionally used to compensate for the potentially biased selection of respondents [**Kish1992**]. Including weights in Bayesian analyses is generally difficult [**Gelman2007**], and, in principle, we should model the data collection process explicitly [**Gelman2013**]. Unfortunately, modelling the data collection process is non-trivial, and we use a weighted pseudo-likelihood instead [**Pfeffermann1996**]. In particular, the likelihood in **??** becomes

$$\tilde{L} = \sum_{j \in U, i \in V_j} w_j \log \rho_{ij} + w_i w_j \sum_{i,j \in U : i \neq j} \log\left(1 - \rho_{ij}\right), \tag{11}$$

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 [**Kish1992**].

Second, the number of terms contributing to the second term of **??** becomes prohibitively large as the number of seeds grows—especially when evaluated repeatedly by a Metropolis-Hastings sampler. Instead, we use a random subset of seed pairs to evaluate the log likelihood. In fact, the number of controls does not have a significant impact on the inference as long as the number is sufficiently large, as discussed in **????**.

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. In 2004, respondents were asked about the demographic background of people "with whom they discuss important matters", which tends to elicit close ties [**Marin2004**]. 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 **??**. For

Variable	Seed coding	Nominee coding	$f(x, y)$
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 1: *Coding of the demographic variables for the General Social Survey together with the feature maps for each variable.*

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 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.*

example, 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.

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 individuals with one or more missing attributes, as shown in **??**. Such a complete-case analysis can introduce biases if the data are not missing completely at random [**Rubin1976**], but handling the missing data in a principled fashion would, once again, require us to develop a model for demographic attributes [**Pigott2001**, **Gelman2013**].

For each demographic attribute, we define a feature for the logistic kernel in **??**, as listed in the last column of **??**. To standardise the features $f(x_i, x_j)$, we subtract their mean and divide non-binary features by twice their standard deviation [**Gelman2008a**]; 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 [**Gelman2008a**] 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 **Gelman2008**, 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

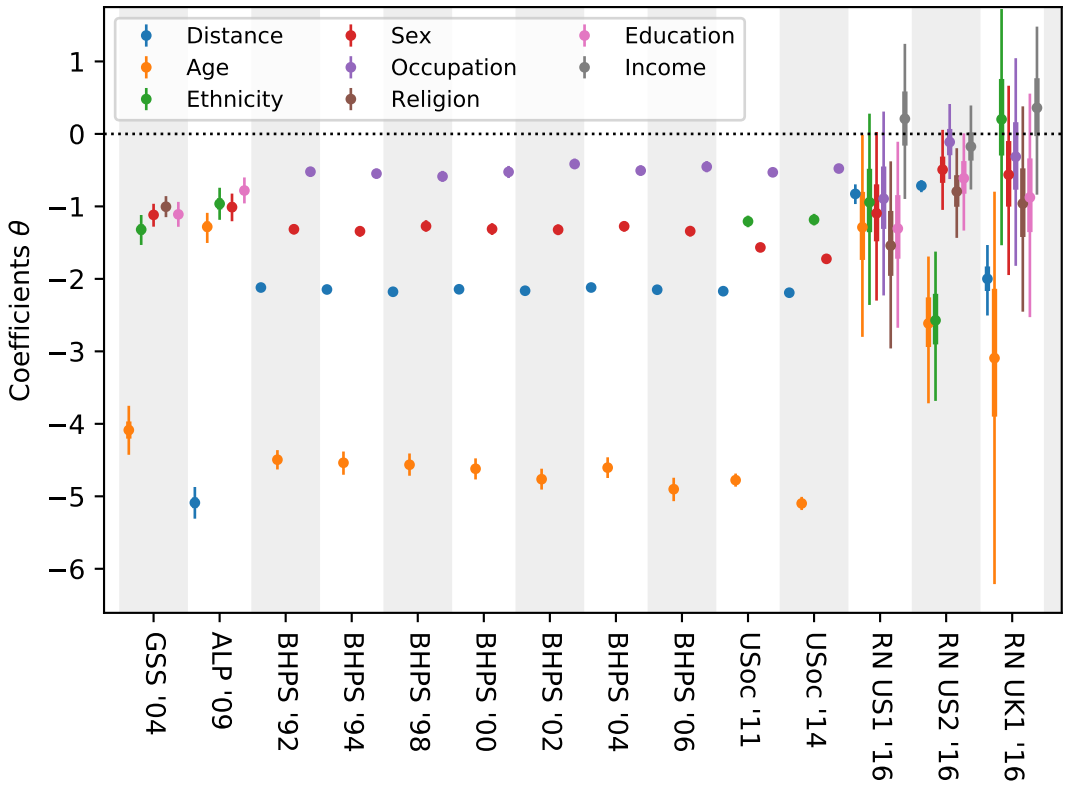

Figure 2: *Age and physical separation have a strong impact on connection probabilities.* Logistic kernel parameters inferred from ego network data for each dataset are shown. Markers represent the posterior mean, thick error bars represent the interquartile range, thin error bars represent the 95% posterior interval.

parameters. Specifying the scale parameters would have been more cumbersome without feature standardisation. For example, different scale parameters need to be chosen for the parameter controlling age homophily depending on whether age is recorded in months, years or decades. We set $\alpha_1 = 10$ because the bias parameter could change significantly depending on the number of controls used [**Gelman2013**].

The inference is run as described in **??**, and the resulting parameter estimates are shown in the first column of **??**. 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.

Hipp2009 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 homophily on age is strong for friendships [**McPherson2001**] (although not as strong as spatial homophily), age is available for most datasets, and is typically uniformly coded in years. In contrast, physical separation is often not available or coded

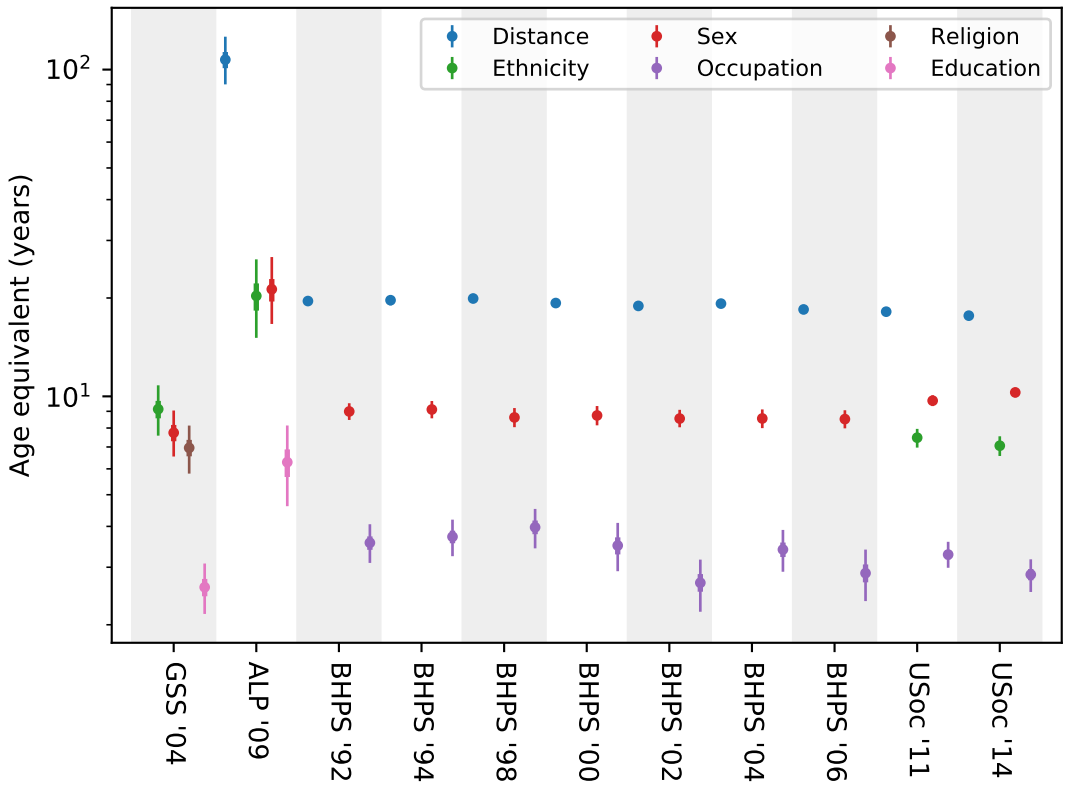

Figure 3: *Converting regression parameters into age equivalents makes the comparison of different features more intuitive.* For binary features (sex, occupation, religion, ethnicity, and distance for the American Life Panel), the equivalent number of years corresponds to a change from having the same attribute to having a different attribute. For non-binary features (education, distance), the equivalent number of years corresponds to a unit increase in the feature. Markers represent the posterior mean, thick error bars represent the interquartile range, thin error bars represent the 95% posterior interval. Age equivalents for the American Life Panel are overestimated (see **??** for details).

heterogeneously across different datasets. For example, the American Life Panel only provides location information at the state level to protect the privacy of respondents (see **??**), whilst the British Household Panel Survey recorded distance between seeds and nominees as ordinal data (see **??**). 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 **??**, corresponds to three years, as shown in the first column of **??**.

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" [**Mihaly2009**]. We only include nominees who are friends of seeds and exclude kinship ties.

We aggregate the ages and educational attainments of seeds to match the coarser coding of nominees, as shown in **??**. The joint effect of ethnic differences and whether people identify as Hispanic is still unclear [**Smith2017a**]; 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 could not identify with any of the provided choices for ethnic identity and selected "other" compared with $< 1\%$ for respondents who did not identify as Hispanic.

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 a result of regression dilution caused by measuring ages imprecisely [**Hutcheon2010**]. Consequently, the age equivalents in the second column of **??** are overestimated. 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 [**Institute-for-Social-and-Economic-Research2000**,

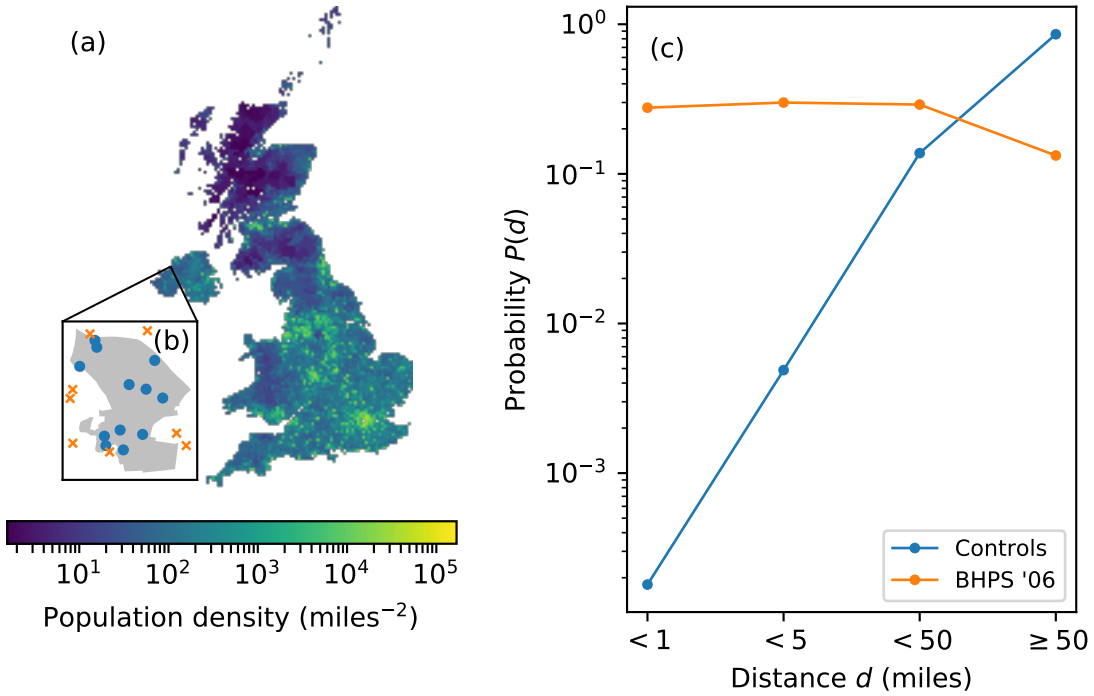

Figure 4: *We obtain the distribution of distances between people using Monte Carlo sampling.* Panel (a) shows the population density in the UK derived from lower layer super output areas. Panel (b) illustrates the sampling method for home locations using rejection sampling. The reference distribution of distances between randomly chosen home locations and the distribution of distances between seeds and nominees is shown in panel (c).

Institute-for-Social-and-Economic-Research2000]. 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.

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 for the likelihood in **??**. But data on the residential location of seeds is not made available to protect their privacy such that we cannot evaluate control features for physical separation. Fortunately, we can sample the home locations of respondents[1] using population estimates and the geographic boundaries of

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

lower layer super output areas (LSOAs). LSOAs are census reporting areas and have a few thousand inhabitants each [**Lsoas**]. We approximate the distribution of distances between residents of the UK using rejection sampling, as illustrated in **??**: 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.

In the Understanding Society survey, 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, $f_{\text{ethnicity}}(x_i, x_j) = 0.5$.

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 [**Hoffmann2016**]. Similarly, we omitted data from the BHPS in 1996 because physical separation between friends was not recorded. As shown in **??**, 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 [**Hoffmann2017**].

3.5 Considerations for future surveys

Having considered data from nine large surveys as well as collecting data ourselves using an online survey panel provider, we reflect on how to improve data collection approaches to learn about Blau space. On the one hand, asking seeds about the demographic background of their social ties can result in noisy Blau space coordinates: seeds may not have accurate information about nominees, or nominees may deliberately portray themselves

inaccurately. For example, men may exaggerate their height on dating platforms because women tend to prefer taller men [**Bruch2016**]. On the other hand, recruiting nominees to provide information about their own Blau space coordinates is resource intensive or, in the case of the surveys we conducted, suffers from small response rates. There is no silver bullet. Furthermore, questions eliciting social ties should be asked first such as not to prompt seeds to think of nominees in the context of a previous question.

Rather than relying on a simple invitation from a friend, future surveys should devise methods to explicitly incentivise nominees to participate in a follow-up survey. For example, qualitative studies of hard-to-reach populations using snowball sampling usually offer an incentive to both the seed and the nominee [**Biernacki1981**]. Unfortunately, even if appropriate incentives have been identified, capturing the attention of social media users remains a challenge.

Alternatively, questionnaires could be modified to record more accurate information about nominees collected from seeds. For example, a respondent who reports the age of one of their friends as an integer multiple of ten could be asked a follow-up question to determine whether the previous answer was an exact age or the respondent's best guess. Rather than asking respondents about potentially sensitive information such as income, proxy information that is more readily available could be collected [**Po2012**]. For example, the number of cars a person owns and whether they own or rent their accommodation could be proxies for wealth or income. Arguably, such proxies may be more relevant for the formation of social ties than income itself because the latter is latent and the former reflect the lifestyle of individuals.

Questions regarding ethnicity should allow respondents to provide multiple answers such that people with mixed ethnic backgrounds can express their identity. In fact, they should explicitly be encouraged to do so to reduce the number of free-form responses that need to be recoded manually. In the US, allowing multiple answers may also help to resolve the question of what it means to be Hispanic in the context of friendship formation [**Smith2017a**]. Furthermore, providing a pan-ethnic option "Asian" rather than specific ethnicities may be more appropriate [**Oyserman1997**]. Questions regarding education should include an option for "some college", and questions probing religious affiliation should include specific Christian denominations.

Whilst it may be necessary to use a coarser coding if seeds are asked to report the demographic background of nominees, the same coding should be used for seeds. Otherwise, attributes for seeds have to be laboriously (and sometimes ambiguously) recoded. Whenever possible, aggregation of attributes such as age into bins should be avoided because it limits the ability to infer kernel parameters [**Hutcheon2010**], as we saw for the American Life Panel.

Finally, the kernel parameters should be inferred jointly: studying them independently could paint a dire picture of society. For example, gender differences and income differences are likely correlated because of the pay gap [**Arulampalam2007**]. Assuming relationships are homophilous with respect to both attributes, fitting kernels for gender and income separately would result in larger, more negative kernel parameters than learning the kernel jointly [**Smith2014**]. In other words, society would appear to be more discriminatory than it really is.

3.6 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 pairwise segregation measure to obtain a distance matrix

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

where $\hat{\theta}$ is the posterior mean of the kernel parameters discussed in **??**. We use multidimensional scaling (MDS) to embed the respondents in a two-dimensional space [**Borg1996**], as shown in **??**. 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 [**Hastie2009**]. The second dimension captures sex and ethnicity. As expected from **??**, 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 connect with one another.

Panel (b) of **??** also shows the social isolation ϕ experienced by individuals in the social space as a 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 pairwise measure in **??**, the global measure can be broken down into components

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

because it is a linear functional of φ: each component contributes to the segregation of society. For each of the datasets discussed in **??**, we evaluate an estimate of the contributions to the global segregation measure

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

where the sum is over all pairs of seeds U. As shown in **??**, physical space has by far the most significant impact on how members of society connect with one another. Interestingly, the spatial contributions to segregation derived from the online surveys we conducted in the US are comparable with the online survey in the UK, even though spatial homophily is more pronounced in the UK, as shown in **??**. A possible explanation

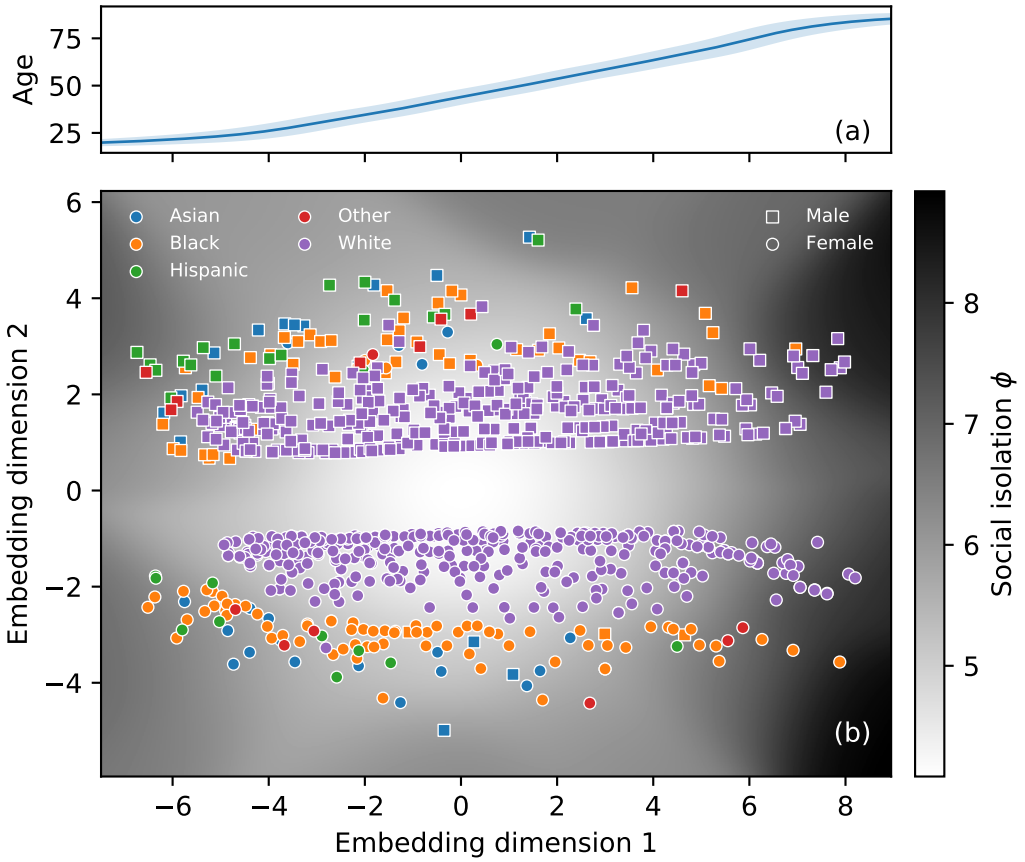

Figure 5: *A lower-dimensional embedding of the inter-node distances reveals an interpretable social space.* In panel (a), the mean and standard deviation of ages are shown as a function of the first embedding dimension as a solid line and a shaded region, respectively. Panel (b) shows a scatter plot of respondents in a two-dimensional embedding space whose coordinates were obtained from the social distance 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.

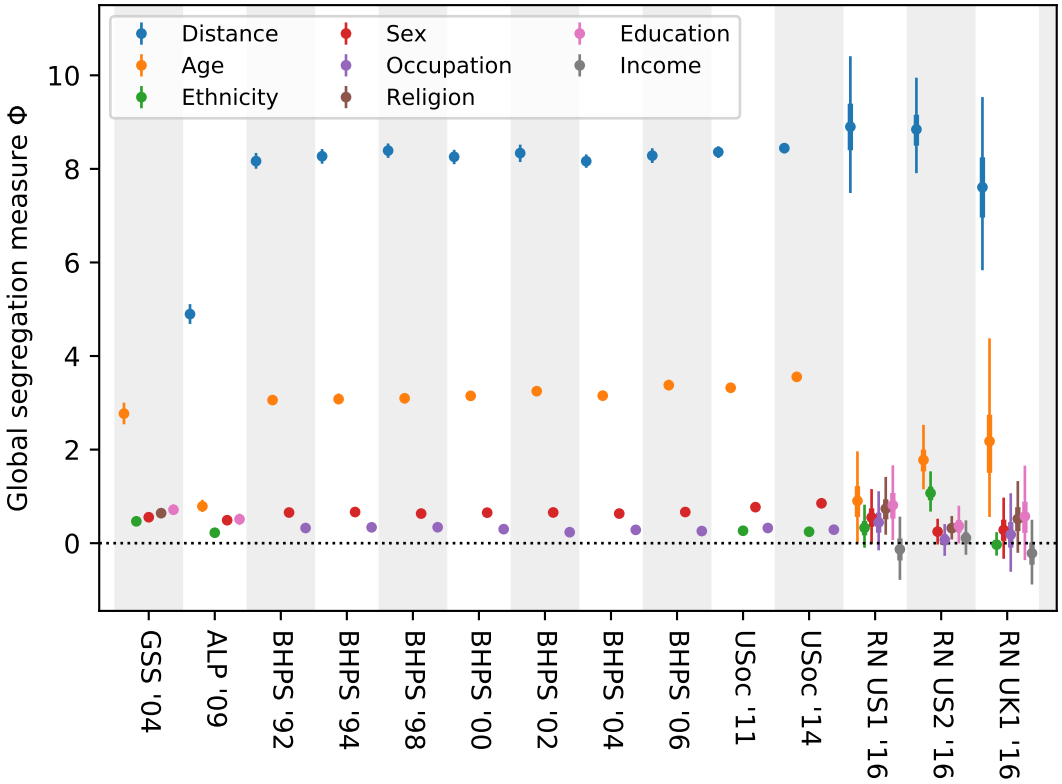

Figure 6: *Physical separation and age differences are the most important factors preventing integration of society for all datasets.* Markers represent the posterior mean of the contributions to the global segregation measure for each feature, thick error bars represent the interquartile range, and thin error bars represent the 95% posterior interval.

for this observation is that individuals adjust their preferences based on the local population density [**Backstrom2010**]. In other words, the number of "intervening opportunities" [**Stouffer1940**, **Illenberger2013**] for making connections is more important than the physical distance such that individuals in low-density areas are less discriminative and form longer-range ties [**Liben-Nowell2005**].

The importance of physical space highlights that our suite of segregation measures does not distinguish between choice homophily and opportunistic homophily [**Franz2010**]. The former is a result of individuals having an active preference to connect with others who are alike, and 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.

The large contribution to the global segregation from spatial effects in the British Household Panel Survey (BHPS) and Understanding Society (USoc) compared with the American Life Panel (ALP) are at odds with the size of the kernel parameters we inferred in **??**. Whilst standardising the features makes the comparison of kernel parameters easier [**Gelman2008a**], correlations amongst the features and skewed feature distributions can make them misleading [**Greenland1986**]. But Φ_l measures the average social distance due to a feature l, and it does not depend on how or whether the corresponding feature was standardised. Thus, the contributions to the global measure are better suited for assessing the relative impact of different dimensions of Blau space.

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 is an exception because of the regression dilution [**Hutcheon2010**] discussed in **??**. Age homophily is known to be particularly strong for friendship networks [**McPherson2001**], and may have had an impact on the recent vote for the UK to leave the European Union and the general election in 2017 [**Goodwin2016**, **Jennings2017**]: we hypothesise that the homogeneity of voting behaviour with respect to age and physical location was in part a result of echo-chamber effects caused by spatial and age homophily. Age homophily and the corresponding segregation measure are weaker in the online surveys we conducted, as discussed in **??**.

Homophily with respect to sex, ethnicity, education, religion, and occupation make similar, small contributions to the segregation of society. Importantly, the global measure captures the average contribution to social isolation and can be small either because there is little homophily or because there is a large majority group, as evident from **??**. 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 have considered a generative model for social networks embedded in Blau space spanned by the demographic attributes of members of the network. We showed that the likelihood for partial observations of the social network obtained using surveys reduces to a weighted case-control likelihood for binary regression. In particular, for logistic connectivity kernels, a standard logistic regression likelihood can be used to infer the kernel parameters.

We inferred the parameters of a friendship connectivity kernel for nine datasets from the UK and two datasets from the US. 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[2]. Our observations, together with a study by **Mossong2008** 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.

We dropped individuals with incomplete demographic attributes from our analysis which may have biased the kernel parameters [**Rubin1976**]. More sophisticated approaches such as model-based imputation may improve the inference [**Pigott2001**].

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 [**Currarini2009**]. However, exponential random graph models may be able to better capture the nature of social networks [**Wimmer2010**]. Furthermore, we have used a connectivity kernel that: (a) is symmetric and cannot identify whether there is a status order in society [**Chan2004**, **Ball2013**], 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 [**Bhattacharya2016**]—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. The former would allow us to measure which people are particularly sociable, and the latter would reveal which demographic backgrounds are particularly attractive to others. Ultimately, learning a connectivity kernel without a pre-specified parametric form should be considered [**Frolich2006**].

Because our approach is based on a generative model of social networks, it can be used to synthesise social networks with realistic attribute correlations amongst connected individuals. Such networks can be shared without risking the privacy of individuals, but it is unclear for which tasks synthetic networks are sufficient. For example, our model will reproduce attribute correlations but is unlikely to be suitable for the study of degree distributions.

[2]The BHPS and USoc are longitudinal studies such that consistent parameter estimates are less surprising.

Finally, our model can be used to impute missing attributes of individuals in social networks. Using Bayes' theorem, the probability distribution over attributes x_i of an individual i is

$$P\left(x_i|A, x_{\{j\in N:j\neq i\}}, \theta\right) \propto P(x_i) \prod_{j\in N:j\neq i} P(A_{ij}, A_{ji}|x_i, x_j, \theta),$$

where $P(x_i)$ is the distribution of attributes in the general population, and the second term accounts for observed connections. In contrast to other methods which generally only produce point estimates [**Wang2013**, **Backstrom2010**], our approach yields a posterior distribution over demographic attributes that captures uncertainties.

In this **??**, we developed a suite of segregation measures that are applicable to multivariate attributes irrespective of their type. The measures describe segregation at different scales: first, the pairwise measure quantifies the social distance between individuals. Second, we can assess how isolated an individual is in society using the individual measure. Finally, the global measure allows us to quantify how segregated society is on average.

The measures are derived from a generative network model facilitating the study of hypothetical societies. Fitting the generative model to data allows us to infer the segregation measures and quantify their uncertainties. Using the survey data discussed in **??**, we establish that spatial separation and age are the most important factors segregating friendship networks.

For certain logistic connectivity kernels, including those used in this thesis, the pairwise measure is a metric for Blau space and allows us to quantify social distance in a principled way. We used the metric to evaluate the social distance amongst respondents of the GSS. We used a lower-dimensional embedding of the respondents to explore Blau space, corroborating our findings that age has a profound impact on restricting friendship formation.

Whilst the global measure is able to summarise segregation at the level of an entire society, it is not suitable for studying the social isolation of minority groups because of its utilitarian foundations. Future work should consider other summary statistics of the pairwise and individual measure in the context of addressing particular social problems. For example, using equal weights for different ethnicities in **??** 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.

A Demographic attributes and feature maps for surveys

Variable	Seed coding	Nominee coding	$f(x, y)$
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 3: *Coding of the demographic variables for the American Life Panel together with the feature maps for each variable.*

Variable	Seed coding	Nominee coding	$f(x, y)$
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) ≥ 50 miles but still in the UK	[b]
Ethnicity??	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 **??**.
[c] See **??** for a detailed description of the feature map.

Table 4: *Coding of the demographic variables for the British Household Panel Survey and Understanding Society together with the feature maps for each variable.*