Latent surface models for networks using Aggregated Relational Data*

Tyler H. McCormick
Department of Statistics
Department of Sociology
University of Washington
Seattle, WA 98103

Tian Zheng
Department of Statistics
Columbia University
New York, NY 10027

Abstract

Despite increased interest across a range of scientific applications in modeling and understanding social network structure, collecting complete network data remains logistically and financially challenging, especially in the social sciences. This paper introduces a latent surface representation of social network structure for partially observed network data. We derive a multivariate measure of expected (latent) distance between an observed actor and unobserved actors with given features. We also draw novel parallels between our work and dependent data in spatial and ecological statistics. We demonstrate the contribution of our model using a random digit-dial telephone survey and a multiyear prospective study of the relationship between network structure and the spread of infectious disease.

The model proposed here is related to previous network models which represents high dimensional structure through a projection to a low-dimensional latent geometric surface—encoding dependence as distance in the space. We develop a latent surface model for cases when complete network data are unavailable. We focus specifically on Aggregated Relational Data (ARD) which measure network structure indirectly

*The authors appreciate the support of the Columbia Applied Statistics Center, the Columbia Population Research Center, and the Statistical and Applied Mathematical Sciences Institute (SAMSI), NSF SES 1023176, USARO 62389-CS-YIP and a Google Faculty Award. We also thank Peter Killworth, Russell Bernard, Chris McCarty, two referees, an Associate Editor, and the Editor.

by asking respondents how many connections they have with members of a certain subpopulation (e.g. How many individuals do you know who are HIV positive?) and are easily added to existing surveys. Instead of conditioning on the (latent) distance between two members of the network, the latent surface model for ARD conditions on the expected distance between a survey respondent and the center of a subpopulation on a latent manifold surface. A spherical latent surface and angular distance across the sphere's surface facilitate tractable computation of this expectation. This model estimates relative homogeneity between groups in the population and variation in the propensity for interaction between respondents and group members. The model also estimates features of groups which are difficult to reach using standard surveys (the homeless, for example).

Keywords: Bayesian methods, density estimation, hard-to-reach groups, partially observed social network

1 Introduction

Social network data consist of relationships (knowing, trusting, etc.) between individual actors, or egos, and another member of the network, known as the alter. Network data are, for example, critical for understanding broad patterns of human behavior (e.g. McPherson et al. (2001)) and for examining the spread of diseases (see Morris (1993), for example). A core statistical challenge associated with social networks is modeling higher order dependence structure. This issue has given rise to a number of statistical models, with one recent attempt being the family of latent geometry models first proposed for networks in Hoff et al. (2002).

Models following from Hoff et al. (2002) assume that the actors in the network form ties independently given their (latent) position in some unobservable "social space." Much like principal components or multidimensional scaling, these models begin with a (likely) high-dimensional feature space and produce a multidimensional geometric representation. The propensity for two individuals to form a tie in the network is inversely related to the

distance between the two in the latent geometry. This latent geometry naturally captures dependence structure in the network. Transitivity (a friend of a friend is likely a friend), for example, is represented through the triangle inequality.

Despite increasing interest in this class of model, such techniques are applicable only when the entire graph is observed. Collecting a complete graph is typically financially and logistically difficult, especially in the social sciences. From a scientific perspective, these data collections issues result in a generalized lack of knowledge about the nature of variation in the day-to-day interactions of individuals (DiPrete et al., 2011; McPherson et al., 2001).

This paper derives a latent surface model for partially observed or sampled network data. We refer to this model as "latent surface" because we use angular (or arc) distance between points on the unit hypersphere to model the propensity for interaction, in contrast to Euclidean distance measured on a hyperplane typically associated with "latent space" models. Our approach begins with a model on the complete graph, then derives the form of the latent surface for the sampled data. The complete graph model is related to the "projection model" described in Hoff et al. (2002). This approach has two key advantages (i) model choices can be made on the complete graph and (ii) the framework yields an explicit relationship between complete graph features and the sampled data, thus illuminating the impact of the sampling procedure. To facilitate interpreting our model in terms of social structure, we also relate structure on the latent surface to overdispersion. Overdispersion describes the variation in relative propensity for a respondent to form ties with members of a particular social group. Zheng et al. (2006) describe overdispersion as an indicator of the likelihood of having exactly one tie to a particular population group, or subopulation.

We focus on data, known as Aggregated Relational Data (ARD), collected through standard surveys using questions of the form "How many X's do you know?". Here, X, represents a subpopulation of interest. ARD are typically used in two situations. First, ARD are used when members of the target population, X, are difficult to enumerate. DiPrete et al. (2011), for example, use ARD to infer properties of respondents' "acquaintanceship" network, where

acquaintances are people the respondent has frequent interactions with but are likely not a close friend. Each respondent's acquaintanceship network has hundreds of individuals and, thus, is not practical for a respondent to enumerate directly. Second, ARD are used to learn about "hard-to-reach" populations. These groups include individuals who are reluctant to report membership in a given group, in many cases because of social pressure or stigma (Shelley et al., 1995). Individuals who have a certain sexual orientation, have a particular political belief, or have some occupations, for example, may not be comfortable revealing this information to an unfamiliar survey enumerator. ARD questions leverage the social networks of survey respondents to reach individuals that are otherwise difficult to access. In this paper, we focus on the second common use of ARD and demonstrate that our proposed method provides new insights into social structure associated with hard-to-reach populations.

We begin in Section 1.1 by presenting our motivation for exploring network structure in ARD and introducing two datasets that we will analyze in subsequent sections. One of these datasets contains only ARD while the other contains individual ties; we use this second data source for validation. Section 1.2 reviews the latent surface model for complete graphs and discusses features that we will use to derive the latent surface model for ARD. Next, Section 2.1 begins with the complete-network model presented in Section 1.2 and derives a latent surface representation of ARD. Section 3 develops a formal latent surface model for ARD, then discusses computation and model fitting issues. Section 4 presents results for this model on two datasets first described in Section 1.1. Section 5 provides a discussion and conclusion.

1.1 ARD and hard-to-reach populations

If respondents could recall perfectly from their network and had full knowledge of all of the group memberships of all other population members, then ARD data would be "equivalent" to asking a respondent if he/she knows each member of a particular group of alters. Rather than reporting these ties individually as in the complete network case, however, our data

consist of only the total number of links the respondent has with Michaels.

ARD surveys typically consist partially of straightforward populations (e.g. people named Michael) used to estimate features of respondents' networks, such as the degree (or personal network size) or rate of mixing between population groups. First names are particularly useful for learning about network structure in the U.S. population since many aggregate features of alters with a given name are available from the Census Bureau and Social Security Administration. Reference groups can also include individuals with a certain occupation or living in a certain geographic region. These names mitigate difficulty of interpreting the axes of the latent surface. Individuals named Robert, for example, are mostly older males. Additional groups, such as hard-to-reach groups, are the primary groups of interest. There is typically little information available about the social networks of these individuals. The network positions and characteristics of populations with unknown demographic make-up (those who are homeless, for example) can then be interpreted in reference to the position of the population with known characteristics.

We focus on the use of ARD to learn about hard-to-reach populations. Previous work used ARD to estimate limited features of these populations. UNAIDS,the joint United Nations program on HIV/AIDS, for example, currently uses ARD (among other techniques) to estimate the sizes of populations most at-risk for HIV/AIDS (UNAIDS, 2003). Shelley et al. (1995) uses ARD to explore social isolation, finding that individuals with HIV had smaller networks than average. With smaller networks comes fewer ties to other members of society, making these individuals more isolated and further ostracized. McCormick and Zheng (2012) use ARD to estimate the demographic profiles of hard-to-reach populations (e.g. the proportion of males under 20 that are clients of commercial sex workers). These methods require detailed information about the demographic distribution of several populations.

In this paper, we demonstrate our method using two datasets. First, McCarty et al. (2001) asked ARD questions on a random-digit dial telephone survey of the U.S. population. These data consist of a series of questions including reference populations (e.g. people with

a particular first name) and several populations where little is known about their structure (e.g. individuals who are HIV positive or homeless individuals). Our second data example comes from a prospective study examining the relationship between network structure and infectious disease propagation, known as Project 90 (Klovdahl et al., 1994; Rothenberg et al., 1995; Woodhouse et al., 1994; Morris, 2004). Unlike the McCarty et al. (2001), the Project 90 (also referred to as the Colorado Springs Study) data contain information about ties between specific individuals in the network. These data consist of a graph with about 6,000 nodes collected by a research team attempting to gather a network census of heterosexual individuals at high risk for contracting HIV (particularly drug injectors and sex workers and their drug/sex partners). We use the Project 90 graph to simulate ARD on a population consisting of many hard-to-reach individuals, then evaluate the ability of our method and others to recover (known) properties of the graph.

1.2 Latent surface models for complete graphs

We begin by reviewing the latent surface model for the completely observed network data then derive the latent surface model for ARD. First, consider two actors i and j whose relationship is described by the $n \times n$ sociomatrix Δ where $\delta_{ij} = 1$ if there is a link between i and j, and 0 otherwise. We refer to i in this example as the *ego* and j as the *alter*. The propensity to form ties ($P(\delta_{ij} = 1)$) between two actors i and j is proportional to the distance between i and j in the latent geometry. Further, these propensities are conditionally independent given the latent distance between i and j (see Hoff et al. (2002) for a full discussion). In the generalized linear model framework, these conditions result in the following model formulation, akin to the "projection model" in Hoff et al. (2002):

$$\begin{aligned} \xi_{ij} &= g_i + g_j + \zeta \mathbf{z}_i' \mathbf{z}_j \\ P(\delta_{ij} = 1 | \xi_{ij}) &= \mathrm{h}(\xi_{ij}) \end{aligned}$$

where $h(\cdot)$ is the link function, $\mathbf{z}_i$ and $\mathbf{z}_j$ are vectors giving the positions on the latent surface of i and j, and ζ is a coefficient scaling the overall influence of the latent component. The g_i and g_j terms represent gregariousness. Gregariousness is an individual random effect that describes the overall propensity for an actor to form ties, or popularity. Gregariousness is related to (though not equivalent to) an actor's network size. Under this model the influence of the latent component represents additional variability explained by social structure in excess of a null model where the propensity for i and j to form a tie depends only on the popularity of i and j.

2 Latent surface representation of overdispersion

In this section we derive a latent surface representation for ARD. We begin by considering the two main parts of our model. First, in Section 2.1 we derive a likelihood by formalizing the concept of aggregation across unobserved alter groups. Second, Section 2.2 describes our model choices about the latent geometry and provides a discussion of the implications of those choices for modeling network structure. In Section 2.3 we show that we can combine the two model components described in Sections 2.1 and 2.2 to formulate a computationally tractable likelihood. Section 2.4 compares our latent surface representation of overdispersion with a related approach to quantifying uncertainty in ARD presented by Zheng et al. (2006).

2.1 Aggregation across alter groups

As described in Section 1, a simple conceptualization of "How many X's do you know?" data involves asking respondents if they know every member of a set of subpopulations, then reporting only the aggregate number known in that subpopulation. We now mathematically formalize the notion of aggregation and derive a latent surface model for ARD. Specifically, for respondent i and subpopulation G_k we observe $y_{ik} = \sum_{j \in G_k} \delta_{ij}$. Conditional on latent positions, $(z_i, z_{j \in G_k})$, the δ_{ij} terms for all i and j are independent Bernoulli random variables,

each with a small probability of success. We can thus model y_{ik} as following (approximately) a Poisson distribution with rate $\lambda_{ik} = \sum_{j \in G_k} \mathrm{P}(\delta_{ij} = 1 | \mathbf{z_i}, \mathbf{z_{j \in G_k}})$ when the number of individuals in the subpopulation, N_k, is large. The key distinction between ARD and the complete network case is that the alters, $j \in G_k$, are unobserved. Without observing these alters, it is not possible to estimate the complete set of alter latent positions and, thus, the Poisson rate described above. Instead, we approximate the rate by taking the expectation over the latent positions of individuals in group G_k. Specifically, we propose the approximation:

$$\lambda_{ik} \approx N_k \int_{\mathbf{z_{j \in G_k}}} \mathrm{P}(\delta_{ij} = 1 | \mathbf{z_i}, \mathbf{z_{j \in G_k}}) \mathrm{P}(\mathbf{z_{j \in G_k}}) d\mathbf{z_{j \in G_k}}. \tag{1}$$

The approximation in (1) has two key features. First, it means that the model is no longer conditional on the distance between two individuals on the latent surface but now conditions on the *expected* distance between a respondent and a subpopulation. Second, mathematically it introduces distributions on the alters on the latent surface, $\mathrm{P}(z_{j \in G_k})$, with integration over the surface of the latent manifold.

The integration in (1) is conceptually similar to techniques found in population studies in the ecology and spatial statistics literature. In Barber and Gelfand (2007), for example, a researcher had intensity measurements for animal sitings across a series of non-equally spaced points in a predetermined geographic region. The statistical goal was to estimate the density of animals across the entire region based on the observable (fixed) intensity measurements. To relate this situation to the latent surface framework, consider the case where we have only one group on the latent surface. Given the positions of observed respondents, $\mathbf{z_i}$, the number of alters each ego knows in the group of interest, G_k, represents an intensity measurement. We use this intensity measurement, as in the spatial statistics and ecology literature, to estimate a continuous density for the group of interest across the (latent) manifold. In our context, however, we estimate densities corresponding to multiple populations. More importantly, measurement locations are also random and need to be estimated.

The implications of the additional uncertainty from estimating ego latent positions is both substantive and computational. Substantively, estimating ego latent positions facilitates representing network dependence structure parsimoniously. That is, the latent position of and individual with a given intensity measurement can be adjusted to best represent the association between an individual and a group of subpopulations. Computationally, the need to estimate both densities and ego latent positions is a major challenge. The problem is related to work in polynomial approximations (see for example Smyth (1998)) and is feasible with fixed location measurements. Barber and Gelfand (2007), for example, exploit this parallel and use quadrature to approximate the intensity function. In this case, however, the additional uncertainty about respondent latent positions makes numerical techniques extremely burdensome. The model we propose affords a computationally tractable form of the likelihood which avoids this issue.

2.2 Latent surface representation

In this section we describe and discuss our model choices for the latent geometry. To begin, let $\mathcal{S}^p$ be the p dimensional hypersphere. We model the latent positions of respondents, $\mathbf{z}_i$ and $\mathbf{z}_j$, lying on the surface of $\mathcal{S}^{p+1}$, corresponding to a p dimensional latent surface on the $p+1$ dimensional hypersphere. We measure distance using the angular distance corresponding to the length of the arc across the surface of the hypersphere.

Envisioning the latent geometry as a hypersphere has advantages. For example, the hypersphere's self-closure property means that we can assume that the respondent latent position vectors, $\mathbf{z}_i$, arise from a uniform distribution across the sphere's surface. This assumption matches our data where respondents are drawn as random sample from the population and, thus, could take any position in the social network. The hypersphere also has finite surface area. This feature imposes a maximum distance between any two points and implies a lower bound on the relative propensity for two individuals to interact. Since our subpopulations are all relatively large, we expect a number of nonzero counts from each

respondent on some population groups. A lower bound on the relative propensity enforces a type of regularization that prevents the gregariousness parameters from exploding. Hoff et al. (2002) also employed a related strategy that normalizes the latent coordinates so that ζ can be identifiable. Also, the framework is computationally appealing and essential for making the likelihood tractable, as discussed in the following section.

Our choice of latent geometry also presents challenges. First, distance on the hypersphere is more difficult to interpret than Euclidean distance on the plane. From a data visualization perspective, the contours of the sphere make it difficult to calibrate equivalent distances across different parts of the sphere. Next, the finite surface area of the sphere can create difficulties in representing multiple populations. The same property that is appealing when imposing a maximum distance between points can also create difficulties when trying to represent multiple points that are similar to some but dissimilar to others. This issues is partially mitigated by the ability to increase the dimensionally of the hypersphere, though the potential for issues increases as the number of subpopulations increases.

Moving now to the specific parameterization of the space, we do not observe alters j belonging to a subpopulation G_k directly and thus wish to parameterize the distribution of these alters across the latent surface. We assume a von-Mises Fisher distribution over the surface of the latent hypersphere. Dhillon and Sra (2003) call the von-Mises Fisher distribution the "natural distribution" for the directional data. The von-Mises Fisher distribution has the form $f_{p+1}(z; \mu_z, \kappa_z) = C_{p+1}(\kappa_z) exp(\kappa \mu_k' z)$, where the normalizing constant depends on the modified Bessel function and simplifies to $\kappa/2\pi(e^{\kappa} - e^{-\kappa})$ when $p = 2$. The vector of length $p + 1$, μ_z, is the mean direction (not expected value) and the scalar $\kappa_z (> 0)$ is concentration, with circular contour lines. Mardia (1972) and Mardia and Jupp (2000) provide a more detailed description of the von-Mises Fisher distribution and describe its geometric relation with the multivariate Normal distributions. The von-Mises Fisher distribution is symmetric and has a single area of concentration. Additional information about this distribution is presented in the Online Supplement.

2.2.1 Relation with other latent space approaches

Related work (e.g. Braun and Bonfrer (2011)) also parameterizes a spherical latent manifold. In our work, the actual geometric representation of the latent surface is a hypersphere. In Braun and Bonfrer (2011), in contrast, the latent geometry is a hyperplane that is then represented using polar coordinates. This approach can be thought of as using latent hyperspheres with varying radii. Though also computationally appealing in their application, using the Braun and Bonfrer (2011) in our context would require marginalizing over the (latent) radius associated with each alter. That is, the integration in Equation 1 using the Braun and Bonfrer (2011) approach is still over the hyperplane.

2.3 Latent surface model for ARD

We now derive the main result, a latent surface model for ARD derived from a model of the complete graph. We begin with a log-linear model on the completely observed graph:

$$\mathrm{E}(\delta_{ij}|g_i, g_j, \zeta, \mathbf{z}_i, \mathbf{z}_j) = \exp(g_i + g_j + \zeta \mathbf{z}_i' \mathbf{z}_j). \tag{2}$$

In (2) we require only that the gregariousness parameters g_i, g_j have finite first and second moments. Specifically, define the overall tendency for actor i to form ties, or gregariousness, as having distribution $P(g_i) = \mathrm{F}(\mu_g, \sigma_g)$. For any member, j, of subpopulation G_k we model $P(g_{j \in G_k}) = \mathrm{F}(\mu_{g_{G_k}}, \sigma_{g_{G_k}})$. The group-specific gregariousness for members of group G_k reflects an association previously noted in the scientific domains where ARD are often used. Politicians or clergy members typically have larger than average networks, for example, while members of some heavily stigmatized or hard-to-reach groups tend to have smaller than average networks (Killworth et al., 1990). Let N_k be the number of members of subpopulation k and N be the total population size. Further, we define the number of individuals in subpopulation k known by respondent i be defined as y_{ik}.

In this notation, our general approach is as follows. First, we begin by computing the

expectation of y_{ik}. Using the conditional independence result from the latent surface model, we then consider $E(y_{ik}) \triangleq \lambda_{ik}$ where λ_{ik} is the rate parameter of a Poisson distribution. Next, we further refine the form of λ_{ik} such that it depends only on terms that are estimable and interpretable based on the respondents. This step entails representing the terms in λ_{ik} as d_i, the degree or total network size of respondent i, and β_k, the fraction of ties in the network that are with group k. The intensity can then be factored into a term consisting of the number known by respondent i, the overall frequency of ties consisting of group members k and a residual term related to the latent geometry.

As described above, we begin by computing the expectation of our observed data, y_{ik}, the number of people known by respondent i in subpopulation k,

$$
\begin{aligned}
\mathrm{E}(y_{ik}) &\triangleq \lambda_{ik} = \sum_{j \in G_k} \mathrm{P}(\delta_{ij} = 1 | \mathbf{z_i}, \mathbf{z_{j \in G_k}}) \\
&\quad \text{and from (1),} \\
&\approx N_k \int_{z_{j \in G_k}} \mathrm{P}(\delta_{ij} = 1 | \mathbf{z_i}, \mathbf{z_{j \in G_k}}) \mathrm{P}(z_{j \in G_k}) dz_{j \in G_k} \\
&\quad \text{substituting from (2),} \\
&= N_k \int_{\mathbf{z}_{j \in G_k}} \exp(g_i + g_{j \in G_k} + \zeta \mathbf{z}_i' \mathbf{z}_{j \in G_k}) P(\mathbf{z}_{j \in G_k}) P(g_{j \in G_k}) d\mathbf{z}_{j \in G_k} dg_{j \in G_k} \\
&= N_k \exp(g_i) \mathrm{E}_k(\exp(g_{j \in G_k})) \int_{\mathbf{z}_{j \in G_k}} \exp(\zeta \mathbf{z}_i' \mathbf{z}_{j \in G_k}) C_{p+1}(\eta_k) \exp(\eta_k \boldsymbol{\upsilon}_k' \mathbf{z}_{j \in G_k}) d\mathbf{z}_{j \in G_k} \\
&= N_k \exp(g_i) \mathrm{E}_k(\exp(g_{j \in G_k})) C_{p+1}(\eta_k) \int_{\mathbf{z}_{j \in G_k}} \exp\left((\zeta \mathbf{z}_i + \eta_k \boldsymbol{\upsilon}_k)' \mathbf{z}_{j \in G_k} \right) d\mathbf{z}_{j \in G_k} \\
&= N_k \exp(g_i) \mathrm{E}_k(\exp(g_{j \in G_k})) C_{p+1}(\eta_k) \\
&\quad \times \int_{\mathbf{z}_{j \in G_k}} \exp\left(\frac{||\zeta \mathbf{z}_i + \eta_k \boldsymbol{\upsilon}_k||}{||\zeta \mathbf{z}_i + \eta_k \boldsymbol{\upsilon}_k||} (\zeta \mathbf{z}_i + \eta_k \boldsymbol{\upsilon}_k)' \mathbf{z}_{j \in G_k} \right) d\mathbf{z}_{j \in G_k}.
\end{aligned}
$$

In this expression $C_{p+1}(\cdot)$ is the normalizing constant of the von-Mises Fisher distribution. Though this constant depends on the ratio of modified Bessel functions, it is easily computed using standard software packages. The integral now contains the kernel of $\mathcal{M}\left(||\zeta \mathbf{z}_i + \eta_k \boldsymbol{\upsilon}_k||, \frac{\zeta \mathbf{z}_i + \eta_k \boldsymbol{\upsilon}_k}{||\zeta \mathbf{z}_i + \eta_k \boldsymbol{\upsilon}_k||} \right)$. Since the latent geometry is a hypersphere, we can add the appropriate normalizing constant and integrate the above von-Mises Fisher distribution

across it's entire support. This simplification yields

$$\lambda_{ik} = N_k \exp(g_i) \mathrm{E}_k(\exp(g_{j \in G_k})) \left(\frac{C_{p+1}(\eta_k)}{C_{p+1}(||\zeta \mathbf{z}_i + \eta_k \boldsymbol{\upsilon}_k||)} \right).$$

The above expression is the basis for the likelihood of a latent surface model for ARD. The form of the expression is not ideal, however, since it remains in terms of the difficult-to-conceptualize quantity of gregariousness. Further, it contains an expectation across alter gregariousness, meaning that one would need to specify a distribution on this quantity to obtain a numerical result. We reparameterize this expression to be in terms of degree (personal network size) and fractional subpopulation size. This reparameterization facilitates both computation and interpretation.

We now present the steps necessary for the reparameterization. We start first with the respondent degree, d_i. Since we do not observe the alters, we define d_i as an expectation:

$$\begin{aligned}
d_i \triangleq \mathrm{E}(\sum_j \delta_{ij}) &= \sum_j \mathrm{E}(\delta_{ij}) \\
&\approx N \int_{\{\mathbf{z}_j, g_j\}} \exp(g_i + g_j + \zeta \mathbf{z}_i' \mathbf{z}_j) P(g_j) P(\mathbf{z}_j) dg_j d\mathbf{z}_j \\
&= N \exp(g_i) \int_{g_j} \exp(g_j) P(g_j) dg_j \int_{\mathbf{z}_j} \exp(\zeta \mathbf{z}_i' \mathbf{z}_j) P(\mathbf{z}_j) d\mathbf{z}_j \\
&= N \exp(g_i) \mathrm{E}(\exp(g_j)) \frac{1}{A_{p+1}} \int_{\mathbf{z}_j} \exp(\zeta \mathbf{z}_i' \mathbf{z}_j) d\mathbf{z}_j
\end{aligned}$$

where A_{p+1} is the surface area of the $p+1$ dimensional unit hypersphere. Since $\exp(\zeta \mathbf{z}_i' \mathbf{z}_j)$ is the kernel of $\mathcal{M}(\zeta, \mathbf{z}_j)$ we have

$$d_i = N \exp(g_i) \mathrm{E}(\exp(g_j)) \frac{1}{A_{p+1}} \frac{1}{C_{p+1}(\zeta)}.$$

and using the fact that the limiting constant $\frac{1}{A_{p+1}} = C_{p+1}(0)$,

$$d_i = N \exp(g_i) \mathrm{E}(\exp(g_j)) \left(\frac{C_{p+1}(0)}{C_{p+1}(\zeta)} \right). \tag{3}$$

We now move to the fractional subpopulation size. We define β_k as the fraction of ties in the network that are made with members of group k:

$$\beta_k \triangleq \frac{\sum_i \sum_{j \in G_k} \delta_{ij}}{\sum_{ij} \delta_{ij}}. \tag{4}$$

Taking each component of (4) in turn we begin by representing the numerator as:

$$\begin{aligned}
\sum_{ij} \delta_{ij} &= N\mathrm{E}(\sum_j \delta_{ij}) = N\mathrm{E}(d_i) \\
&= N \int_i N \exp(g_i)\mathrm{E}(\exp(g_i)) \frac{C_{p+1}(0)}{C_{p+1}(\zeta)} p(g_i) dg_i \\
&= N^2(\mathrm{E}(\exp(g_i))^2 \frac{C_{p+1}(0)}{C_{p+1}(\zeta)}.
\end{aligned}$$

Turning now to the denominator, we provide the result and leave the full derivation for the Online Supplement. The key insight is that, after defining $\sum_i \sum_{j \in G_k} \delta_{ij} = \sum_i N_k \mathrm{E}(\delta_{ij} | j \in G_k)$, rearranging and taking the expectation over i, we can exchange the order of integration and leverage the symmetry of the inner product to write the expression in terms of the kernel of a von-Mises Fisher distribution. Since the expectation is over the entire surface of the hypersphere, this is now the kernel of a probability distribution integrated over its entire support. Replacing the expression with the inverse of the normalizing constant we have:

$$\begin{aligned}
\sum_i \sum_{j \in G_k} \delta_{ij} &= \sum_i N_k \mathrm{E}(\delta_{ij} | j \in G_k) \\
&= \sum_i N_k \int_{j \in G_k} \exp(g_i + g_{j \in G_k} + \zeta \mathbf{z}_i' \mathbf{z}_{j \in G_k}) p(g_{j \in G_k}) p(\mathbf{z}_{j \in G_k}) d\mathbf{z}_{j \in G_k} dg_{j \in G_k} \\
&= N N_k \mathrm{E}(\exp(g_i)) \mathrm{E}_k(\exp(g_{j \in G_k})) \frac{C_{p+1}(0)}{C_{p+1}(\zeta)}.
\end{aligned}$$

Combining the reparameterized numerator and denominator yields a new expression for β_k:

$$\beta_k = \left(\frac{N_k}{N} \right) \left(\frac{\mathrm{E}_k(\exp(g_{j \in G_k}))}{\mathrm{E}(\exp(g_i))} \right).$$

We substitute $\frac{C_{p+1}(0)}{C_{p+1}(\zeta)} d_i \beta_k = N_k \exp(g_i) \mathrm{E}_k(\exp(g_{j \in G_k}))$ to have

$$\lambda_{ik} = d_i \beta_k \left(\frac{C_{p+1}(\zeta) C_{p+1}(\eta_k)}{C_{p+1}(0) C_{p+1}(||\zeta \mathbf{z}_i + \eta_k \boldsymbol{\upsilon}_k||)} \right).$$

Noting that $||\zeta \mathbf{z}_i + \eta_k \boldsymbol{\upsilon}_k|| = \sqrt{\zeta^2 + \eta_k^2 + 2\zeta\eta_k \cos(\theta_{(\mathbf{z}_i, \boldsymbol{\upsilon}_k)})}$ we have

$$\lambda_{ik} = d_i \beta_k \left(\frac{C_{p+1}(\zeta) C_{p+1}(\eta_k)}{C_{p+1}(0) C_{p+1}(\sqrt{\zeta^2 + \eta_k^2 + 2\zeta\eta_k \cos(\theta_{(\mathbf{z}_i, \boldsymbol{\upsilon}_k)})})} \right) \tag{5}$$

where we now use $\theta_{(\mathbf{z}_i, \boldsymbol{\upsilon}_k)}$ to represent the angular distance (arc distance across the hypersphere's surface) between the respondent's position and the center of subpopulation k. The form of λ_{ik} in (5) completes our latent surface representation for ARD.

Returning now to the arguments in Section 2.1, we have that the y_{ik} follows (approximately) a Poisson distribution when N_k is large. Combining these results we have

$$y_{ik} | d_i, \beta_k, \zeta, \eta_k, \theta_{(\mathbf{z}_i, \boldsymbol{\upsilon}_k)} \sim \text{Poisson} \left(d_i \beta_k \left(\frac{C_{p+1}(\zeta) C_{p+1}(\eta_k)}{C_{p+1}(0) C_{p+1}(\sqrt{\zeta^2 + \eta_k^2 + 2\zeta\eta_k \cos(\theta_{(\mathbf{z}_i, \boldsymbol{\upsilon}_k)})})} \right) \right), \tag{6}$$

which is the likelihood for the latent surface model for ARD. This likelihood is computationally tractable, greatly facilitating model fitting and sampling in practice.

The impact of latent position on the expected number known is weighted by the concentration of a subpopulation, forming a measure similar to a Mahalanobis distance in a Euclidean space. The distance to the center of the subpopulation does matter (akin to to the distance between individuals in the complete-network case), but the impact of this distance is modulated by the concentration of the subpopulation. For diffuse subpopulations, the angular distance between the respondent and the center of the subpopulation has a relatively linear impact on the expected number known, with a gradual increase as the respondent nears the center of the subpopulation. For highly concentrated groups, however, the impact is more extreme. Individuals who are close to the center of the subpopulation are expected to know many members while the expected number known drops precipitously

as distance increases. This feature leads to a relationship between concentration of groups on the latent surface and overdispersion.

2.4 Comparison to overdispersion in Zheng et al. (2006)

To better understand the role of the latent geometry, we compare the model above with the overdispersed model presented in Zheng et al. (2006). The Zheng et al. (2006) model also assumes a Poisson distribution for y_{ik} with $\lambda_{ik} = d_i\beta_k\gamma_{ik}$. Rather than estimating γ_{ik} directly, Zheng et al. (2006) assign a Gamma prior distribution to γ_{ik} with a mean of 1 and shape parameter $1/(\omega_k - 1)$. The γ's are integrated out to yield a Negative Binomial distribution with overdispersion ω_k. This model provides a scalar representation for deviance from random mixing, where the amount of deviation is the same for all group members.

In the latent surface model we can also conceptualize γ_{ik} as controlling the propensity for i to form ties with group k in excess of random mixing, where random mixing refers to the case where $\gamma_{ik} = 1$ and the propensity for ties depends only on a respondent's network size and the size of the alter group. In the latent surface model the form of γ_{ik} remains individual-specific and is given by

$$\gamma_{ik} = \frac{C_{p+1}(\zeta)C_{p+1}(\eta_k)}{C_{p+1}(0)C_{p+1}(\sqrt{\zeta^2 + \eta_k^2 + 2\zeta\eta_k\cos(\theta_{(\mathbf{z}_i,\boldsymbol{\upsilon}_k)})})}. \tag{7}$$

The departure from what would be expected under random mixing now depends on the concentration of the population of interest *and* the distance between the individual and the center of the alter group. Setting $\zeta = 0$ in (7), we see have that $\gamma_{ik} = 1$ and the model simplifies to the "null model" for random mixing presented in Zheng et al. (2006). Further, taking the expectation of λ_{ik} and rearranging the resulting expression yields that γ_{ik} has expectation 1. The variance of γ_{ik} (and therefore overdispersion) increases monotonically as the concentration of the subpopulation η_k increases relative to the general level of the population, ζ. This result can be verified through simulation (not shown).

In summary, the Zheng et al. (2006) qualifies the existence of overdispersion, whereas the latent surface model provides a multivariate representation of how overdispersion varies across groups in the population. After marginalizing over γ_{ik}, Zheng et al. (2006) can still makes statements about the general degree of departure from random mixing for a particular group. With the latent surface model, however, the goal is to understand how those departures from random mixing vary between egos and across alter groups.

3 Latent surface model and computation

In this section we describe model fitting for our latent surface model for ARD. We first describe formally our model and prior structure and conclude by presenting our model fitting algorithm. Then, we discuss identifiability and suggest strategies for future researchers who encounter identifiability issues when fitting similar models.

3.1 Priors and posterior

We give d_i and β_k normal priors on the log scale. As described in the previous section $\mathbf{z}_i$ has a uniform prior distribution across the hypersphere. Respondents in subpopulation k have latent subpopulations with priors $\mathbf{z}_{j\in G_k} \sim \mathcal{M}(\boldsymbol{\upsilon}_k, \eta_k)$. We assume $\boldsymbol{\upsilon}_k$ has a uniform prior across the sphere. We propose Gamma priors for ζ and η_k with conjugate priors on the hyperparameters.

If $\lambda_{ik} = d_i\beta_k \left(\frac{C_{p+1}(\zeta)C_{p+1}(\eta_k)}{C_{p+1}(0)C_{p+1}(\sqrt{\zeta^2+\eta_k^2+2\zeta\eta_k\cos(\theta_{(\mathbf{z}_i,\boldsymbol{\upsilon}_k)})})} \right)$ and θ is shorthand for all parameters, then the posterior is:

$$\begin{aligned}\theta|y \;\propto\; & \prod_{k=1}^{K}\prod_{i=1}^{N}\exp(-\lambda_{ik})\lambda_{ik}^{y_{ik}}\prod_{i=1}^{N}\text{Normal}(\log(d_i)|\mu_d,\sigma_d^2)\prod_{i=1}^{N}\text{Normal}(\log(d_i)|\mu_d,\sigma_d^2) \\ & \times\prod_{k=1}^{K}\text{Normal}(\log(\beta_k)|\mu_\beta,\sigma_\beta^2)\prod_{k=1}^{K}\text{Normal}(\log(\eta_k)|\mu_{\eta_k,\sigma_{\eta_k}^2})\text{Gamma}(\zeta|\gamma_\zeta,\psi_\zeta).\end{aligned}$$

Since it is not easy to sample from the posterior directly, we use Markov-chain Monte Carlo

(MCMC). Before describing our computation procedure, we first introduce identifiability issues associated with the proposed model.

3.2 Identifiability

In this section we describe restrictions on the latent surface and model parameters necessary to ensure identifiability. The most challenging identifiability issue in our model arises because the likelihood depends on the latent positions only through their inner product. We could, for example, rotate the entire representation by a given angle and obtain exactly the same value of the likelihood, creating challenges for computation and interpretability. We address these issues by fixing the centers of a subset of our target subpopulations. This approach is similar to methods used in previous work on latent geometry models. Hoff (2005) for example estimate an initial set of latent positions and then rotate back to configuration most similar to this original orientation at each MCMC iteration. In contrast to previous approaches, where the initial configuration is arbitrary, the choice of where to fix our population centers impacts our visual representation of the structure of the network.

We now discuss the number of fixed population centers and their locations. Throughout our description, we refer to simulation studies using the Project 90 data which are available in the Online Supplement. For our simulation, we fit the latent surface model using a two dimensional latent surface (a three dimensional sphere). We have four groups with some known information and six groups with unknown characteristics representing hard-to-reach groups. We first vary the number of fixed populations from zero to four. Then, we evaluate different locations for a setting with the same number of fixed populations.

There are two primary considerations in deciding how to fix the population centers. First, the number fixed populations has computational implications. Fixing too few populations leaves the latent distances invariant to certain types of rotations. The sampler can then visit equivalent, or nearly equivalent, representations with different parameter values. This results in an overestimate of variability and increased time to convergence. In our simulation studies,

for example, we see lower effective sample sizes for the same number of iterations when we fix zero or one population than when we fix three or four. Fixing too many populations, in contrast, overly constrains the space and can mask important network features.

A second consideration concerns how the fixed populations contribute to the ability to use the latent surface representation to learn about populations with unknown characteristics. Though not necessary for the computational considerations mentioned previously, the fixed populations should ideally be those with some known characteristics. The McCarty et al. (2001) data, for example, contain populations defined by individuals with a certain first name. Information about the age distribution is available from the U.S. Social Security Administration. The characteristics of other groups with unknown characteristics can then be inferred from their positions relative to fixed populations. Separating the fixed positions of groups known to have different characteristics, therefore, also encourages separation in the inferred positions. In the following section, for example, we present results which position the group of individuals who came down with AIDS in the past year close to the fixed center of the group named Christopher. Christopher is a name which is most common among younger males. AIDS is also known to be most commonly found among young males. In this way, we indirectly use prior information about members of groups with known demographic characteristics to translate the geometry of the latent surface into a graphical representation of social structure in the network. Similarly, in our simulation studies with the Project 90 data presented in the Online Supplement, we define populations based on respondents' age and race. These groups are unlikely to be feasible in practice because the number of alters known in each group is likely quite large. Social interactions are often highly divided based on race, however, providing a strong signal of network structure. To demonstrate the impact of the layout of fixed populations, the Online Supplement contains simulation studies where we alter the positions of the center of the same latent groups. The simulations demonstrate that, consistent with previous work on latent space models, multiple orientations are empirically equivalent so long as they preserve a common matrix of latent distances.

Along with being populations with some known demographic information, we also suggest that the fixed populations be heterogeneous and, to the extent possible, uncorrelated. Choosing diverse populations with fixed positions ensures easy differentiation between the groups each population represents. Fixing two female names of the same age, for example, would make it difficult to distinguish respondents with ties to older females to those with older males. These diverse populations should then be spaced far apart on the latent surface. Appropriate spacing encourages the model to use as much of the space as possible, maximizing the distance between estimated positions.

We also take steps to ensure identifiability of additional model parameters. Since d_i and β_k enter the likelihood only through their product, neither is identifiable without additional restrictions. We address this by constraining the total size of a subset of β_k values based on the known sizes of some populations. A similar strategy was effective in Zheng et al. (2006). The coefficient, ζ which modulates the impact of the latent features is identifiable because of the stipulation that the points lie on the sphere. This coefficient is initially used in the Hoff et al. (2002) paper but drops out in the Handcock et al. (2007) paper because it creates an identifiability issue. Hoff et al. (2002) artificially normalize the inner products, while this feature is present naturally as a byproduct of the geometry of the sphere.

3.3 MCMC algorithm

As previously mentioned, the members of the latent subpopulations are never observed directly. Instead, we make inferences about the expected latent distance from a respondent and a member of G_k using the expression obtained from the integration in the previous section. This expression involves modified Bessel functions, but is easily evaluated and can therefore be used for evaluating proposals in Metropolis steps. This approach is a significant computational savings over a Monte-Carlo approximation to the necessary expectation.

Assume the subpopulations, $k = 1, ..., K$, such that $K \geq p$. We fit the model as:

1. For a subset of the subpopulations, $k^{(s)} = 1, ..., K^{(s)}$, fix $\boldsymbol{v}_k^{(s)}$ for identifiability. Number

of subpopulations to fix depends on the dimension of the latent surface. We will use these fixed positions to rotate the latent surface back to a common orientation at each iteration using a Procrustes transformation (see Hoff et al. (2002) for the details).

2. Repeat to convergence for $m = 1, ..., M$

   (a) For each i, update $\mathbf{z}_i$ using a random walk Metropolis step with proposal $\mathbf{z}_i^* \sim \mathcal{M}(\mathbf{z}_i^{(m-1)}, \text{jumping scale})$. One option is to simulate from these distributions at the same time using Hoff (2009). This method would loop over each k anyway, so we use the algorithm proposed by Wood (1994).

   (b) Update $\boldsymbol{\upsilon}_k$ using a conditionally conjugate Gibbs step (Mardia and El-Atoum, 1976; Guttorp and Lockhart, 1988; Hornik and Grün, 2013).

   (c) Update d_i with a Metropolis step with $\log(d_i^*) \sim \text{N}(\log(d_i)^{(v-1)}, (\text{jumping distribution scale}))$.

   (d) Update β with a Metropolis step with $\log(\beta^*) \sim \text{N}(\log(\beta)^{(v-1)}, (\text{jumping distribution scale}))$.

   (e) Update η_k with a Metropolis step with $\eta_k^* \sim \text{N}(\eta_k^{(v-1)}, (\text{jumping distribution scale}))$.

   (f) Update ζ with a Metropolis step with $\zeta^* \sim \text{N}(\zeta^{(v-1)}, (\text{jumping distribution scale}))$.

   (g) Update μ_β, μ_d, σ_β^2, and σ_d^2 using Gibbs steps (details in Online Supplement).

We not turn attention to implementing this algorithm on our two data sources.

4 Results

In this section we describe results from two data sources. First, we present results using data collected by McCarty et al. (2001) using a random-digit dial telephone survey. The second data, the Project 90 data, arise through a prospective study of disease propagation. The McCarty et al. (2001) data demonstrate the possible insights from our method using an actual ARD data collection. The Project 90 data, in contrast, present and opportunity to examine the method's properties when the underlying network is known. Both data also likely content measurement error, which we discuss further in the conclusion.

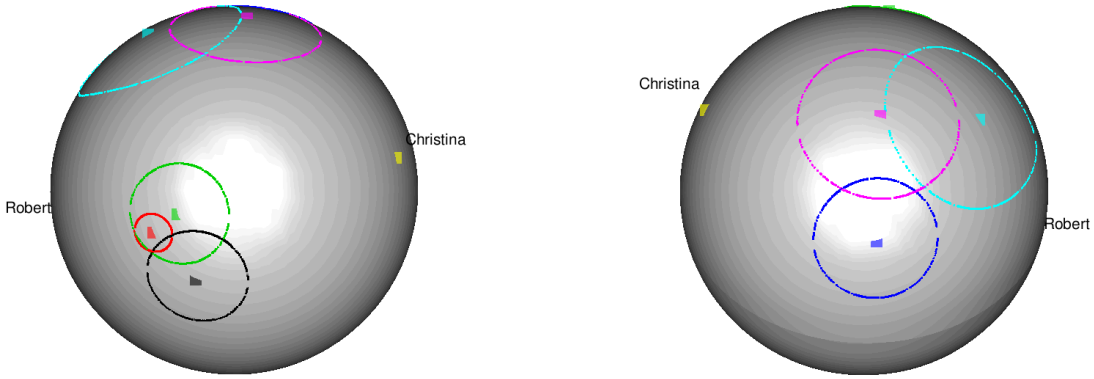

Figure 1: Two views of a spherical latent surface displaying latent positions of six subpopulations: homeless (red), AIDS (black), in prison (green), adopted children (purple), Jaycees (blue), and receiving kidney dialysis (aqua). Individuals in prison, homeless and with AIDS have similar latent positions, though the distribution for homeless individuals is much more concentrated. Dots represent the posterior mode of latent group position and lines represent concentration.

4.1 McCarty et al. (2001) data

We used data from a telephone survey conducted by McCarty et al. (2001) with 1375 respondents. We ran three sampler chains, each with 10,000 iterations, discarding the first half of each chain and keeping every tenth iteration. The remaining chains mixed well with acceptable convergence ($\max \hat{R} \leq 1.1$, see Gelman et al. (2003)) and effective sample sizes ranging from about 300 to about 480. We used a three-dimensional latent surface. We address the sensitivity of our method to the choice of latent dimension in Section 4.3. As noted in Section 3.2, fixing the center of some latent groups facilitates identifiability and interpretation. We fixed the centers of the groups "Robert," "Christina," "Christopher," and "Jacqueline." These names have known demographic profiles from the U.S. Social Security Administration. The population of Roberts is comprised mostly of older males, while the majority of Christinas are younger females. Fixing the same number of male and female names preserves a balance between genders. This encourages diversity in respondent latent positions, since we expect male respondents to know more individuals with male names (and thus have latent positions closer to the fixed male groups), and vice-versa.

Figures 1 and 2 display a latent surface with six subpopulations, with "AIDS" referring to

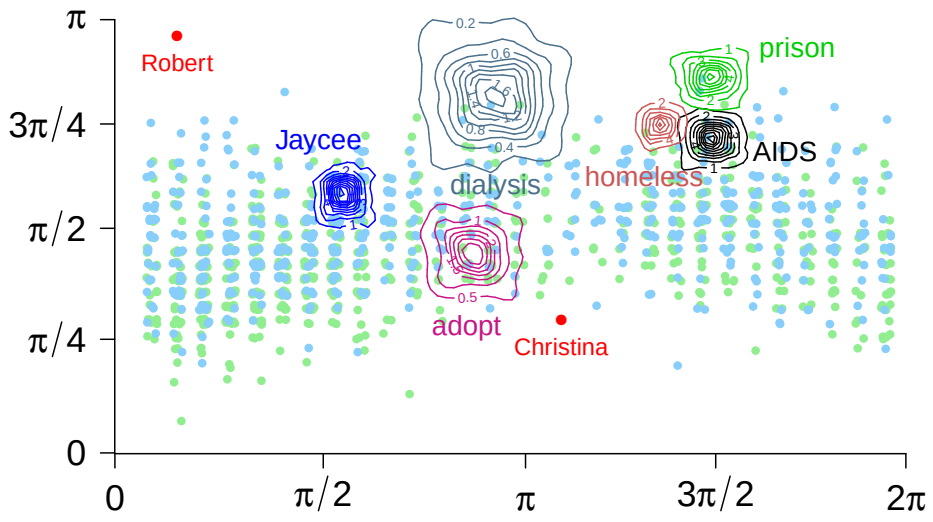

Figure 2: Latent positions of five subpopulations: homeless, AIDS, in prison, adopted children, and receiving kidney dialysis. This figure displays all five groups. Blue dots are the posterior mode of male respondents' latent positions and green dots are females. This figure contains the same information as Figure 1 but has been projected onto the plane. The contours plotted are contours of the posterior predictive distributions of the latent subpopulation members ($p(z_{j\in G_k}|\cdot)$ averaging over the center and concentration on of the subpopulation).

individuals who came down with AIDS in the past year. The contours plotted are contours of the posterior predictive distributions of the latent subpopulation members ($p(z_{j\in G_k}|\cdot)$ averaging over the center and concentration of the subpopulation). The visual representation provided by the latent surface model uses the geometry of the latent surface to represent the dependence structure of the network. The subpopulation of individuals who are homeless has a position similar to those who are incarcerated and those with AIDS but is more concentrated. From this we conclude that the network connectivity of individuals in these three subpopulations are more similar to one-another than they are to individuals who adopt children, but that the subpopulation of homeless individuals is more homogenous.

The concentration of the subpopulations demonstrates the flexibility of the latent surface model in representing various network features. Members of the Jaycees (a civic service

organization for young professionals) and the population of homeless individuals have similar concentrations. This indicates that, for both groups, the propensity of knowing one member of the group is very low. Likewise, they have a comparable overall level of variation in excess of a random mixing model (overdispersion). The latent surface model captures this feature, yet also reflects the diverse nature of individuals who are highly connected with these two groups by placing them on opposite sides of the sphere. Thus, an individual whose latent position is close to the center of the Jaycees will, in expectation, know many members of the Jaycees but is expected to know very few individuals who are homeless.

Comparing with the Zheng et al. (2006) model, the importance of the additional flexibility of the latent surface representation becomes clear. Recall that deviation from random mixing in the Zheng et al. (2006) model is measured using a scalar overdispersion parameter for each population group. This parameter gives a one-dimensional interpretation of the quantity of excess variation. It does not represent, however, the way that this excess structure influences connectivity patterns in the network. Taking Jaycees and homeless individuals again as an example, both Zheng et al. (2006) and the latent surface model estimate very high overdispersion. In Zheng et al. (2006), the overdispersion parameter is nearly identical for the two groups (see Figure 4 in Zheng et al. (2006)). The latent surface model, however, produces similar estimated concentrations in the latent distribution but, as described above, positions the two groups on opposite sides of the latent surface. This distinction can also be captured numerically as the latent surface model corresponds to approximately a thirty percent reduction in root mean-squared error compared to the Zheng et al. (2006) model.

We can also use the latent surface model to infer unknown characteristics of alter groups. If there is a high degree of homophily based on a particular characteristic, then individuals sharing the trait are more likely to interact. In aggregated data, this feature manifests as a higher than expected count of the number of individual a respondent knowns in a particular group. We leverage this feature to provide insights about the composition of alter groups. Figure 3 displays results from regression models where the outcome variable is the

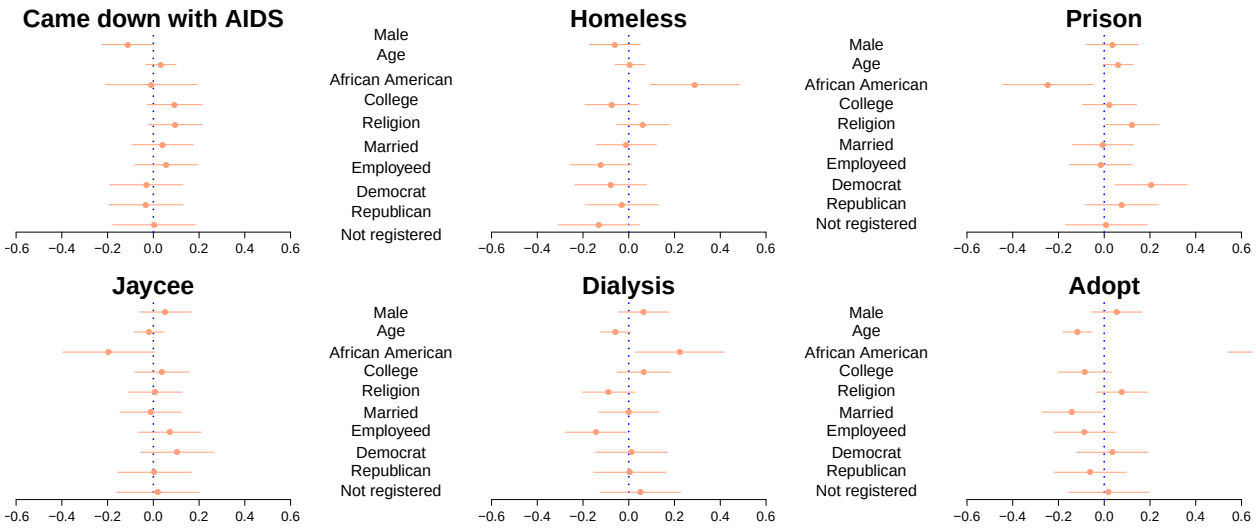

Figure 3: Each graph presents results from a regression model where the outcome is the posterior mode of the individual adjustment to know members of each group (γ_{ik} in (7)). Under an assumption of homophily, this strategy provides information about the composition of alter groups. Not registered indicates that an individual is not registered to vote. Religion indicates the respondent reported belonging to a religious organization. Married indicates the person has never been married.

(standardized) posterior mode individual adjustment over random mixing for members of a particular groups, or γ_{ik} in (7). This term includes both the distance between the individual and the group on the latent surface and the concentration of the alter group. Higher values indicate the individual has an increased propensity to know individuals in the alter group when compared to random mixing. Covariates in the regression are characteristics related to the respondent's demographics and political affiliations.

In cases where there are high degrees of homophily, this approach provides information about the characteristics of hard-to-reach groups. The alter group of homeless individuals has a large coefficient for African Americans. Similarly, the group of individuals adopting children also has a large positive coefficient for individuals who are African American. Examining data from the U.S. Department of Health and Human Services (USDHHS, 2013), nearly one-third of adoptive parents are African American, nearly triple the number expected if adoptions were distributed equally across all races. Similarly, the coefficient for the indicator of attending college is large and negative for homeless individuals but positive for the Jaycees.

Several challenges arise in using these results. The first challenge, akin to sensitivity, is that the quality of results vary based on the degree of homophily. Cases with strong homophily, such as race, will likely provide superior results to situations with less homophily. If politically active individuals are not more likely to interact with one another, for example, then the social network will provide little leverage. Recent work in this area by DiPrete et al. (2011) provides some insights, indicating that employment status, religious behavior, and political opinions produce perceived segregation on nearly the same scale as race. A second potential challenge, related to specificity, is present if individuals in a given group are not the only members of society who are highly connected to that group. Taking again the case of religion and individuals who came down with AIDS in the past year, for example, the strong positive coefficient in Figure 3 indicates that individuals who interact frequently with those who came down with AIDS are likely to report a religious affiliation. This result could indicate that who recently came with AIDS are likely to be active in religious organizations. Alternatively, the individuals whose latent positions are closest to the population of individuals who came down with AIDS could be working as medical professionals or social workers and interact frequently with individuals living with AIDS, particularly shortly after diagnosis. This is particularly likely since the focus is on individuals whose condition has progressed from HIV to AIDS and thus are likely interacting more frequently with the healthcare system. These service professionals may, in turn, be more likely to be affiliated with religious organizations. Finally, the sample of individuals in each of the covariate categories should be representative of the population overall. If for example, the African Americans who responded to a telephone survey were substantially different than other African Americans, then it would be difficult to generalize these results. There is evidence in the data that this may be the case. About half of the African Americans in the survey, for example, reported having gone to college, a fraction substantially higher than the general population. This phenomenon could be responsible for seemingly surprising results, such as the negative coefficient for African Americans in the prison alter group.

4.2 Project 90

As described in Section 1, the Project 90 data consist of a graph with about 6,000 nodes, which we take to be a completely observed graph. The Project 90 data also contains several hard-to-reach populations, similar to situations where ARD are used in practice.

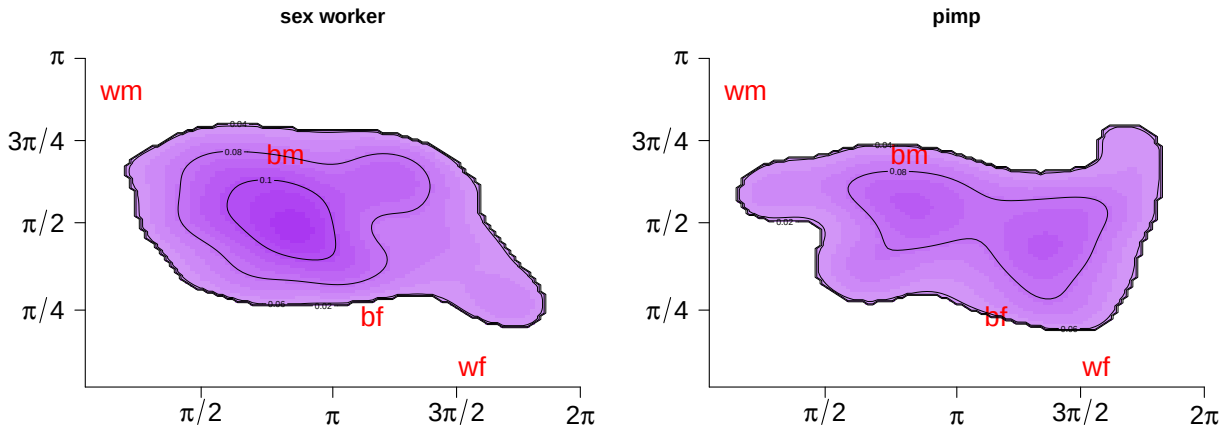

Figure 4: Latent surface for the Project 90 data. Contours of the posterior predictive distribution for sex workers and pimps. Populations with a concentration of females in the complete graph, such as sex workers, have density closer to female reference populations.

Using the graph from the Project 90 data, we simulate ARD sampling and attempt to recover the structure in the complete graph. We begin by taking a simple random sample of respondents in the Project 90 network. For each respondent, we use the respondent's local network to construct ARD by summing the number of connections with members of each group. We used ten groups: four reference groups (Caucasian males, Caucasian females, African American males, and African American Females) and six groups of interest (Drug dealers, Drug cooks, Commercial sex workers, Sex worker clients, disabled individuals, and housewives). We chose the last two groups to compare groups associated with a higher risk of HIV with groups that are more widely distributed throughout the population.

After simulating ARD, we compare the results from the latent surface model with the network structure in the complete graph. First, we evaluate the latent surface model's ability

to capture preferential mixing between population groups. Figure 4 displays the posterior predictive distributions for two populations from the project 90 data. The population of sex workers is concentrated to black males. Looking to the complete graph, sex workers knew about 60% more black males than expected under random mixing. In contrast, sex workers knew about 15% fewer white females than expected. Turning now to pimps, in the complete graph pimps knew about 2.5 times as many drug dealers and about 6 times as many sex workers as would be expected under random mixing. Both sex workers and drug dealers have positions close to pimps on the latent surface. The population of individuals reporting to be housewives, in contrast, has a latent position that is, on average, about five times farther from the estimated latent position for pimps. Comparing to the complete graph, pimps knew about 38% fewer housewives than would be expected under random mixing. We also present a direct comparison of our approach to fitting a latent space model to the complete graph in the Online Supplement. Understanding the tendency for these hard-to-reach groups to intact with others in the population has potential benefits for, among other things, understanding disease transmission or providing targeted services to members of these groups.

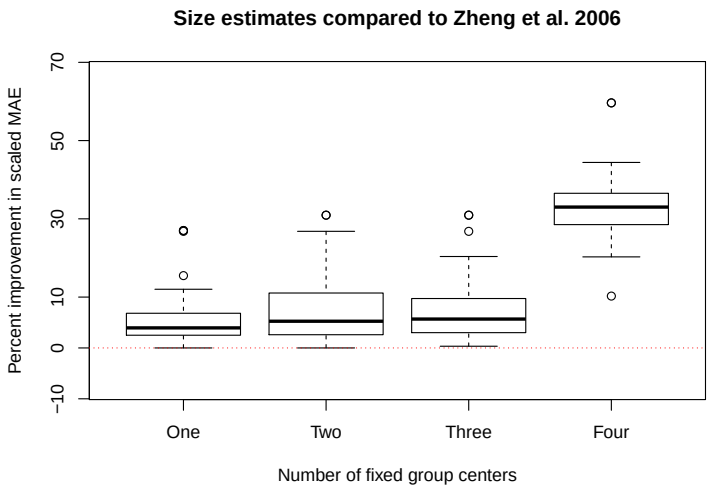

Figure 5: Boxplots of improvement in parameter recovery experiment. Each box presents the results from multiple simulate datasets from the Project 90 network. The height of the graph represents the percent reduction in scaled mean absolute error when comparing the latent surface model with the Zheng et al. (2006) model.

A second comparison involves population size estimation. As previously mentioned, estimating the sizes of hard-to-reach populations is critically important for public health and epidemiology. We perform a simulation experiment using 50 simulated samples of size 750 from the Project 90 network. Once we have simulated ARD, we fit both the Zheng et al. and the latent surface model and compare estimates of network size. In both models, respondent degree also has to be estimated. The four reference groups mentioned above comprise about ninety-six percent of respondents in the network. Respondent's total number known across all four groups, therefore, is effectively the respondent's degree. This feature removes variation caused by error in estimating respondent degree isolates any error to the population sizes. Figure 5 presents the results fixing the locations of between one and four of the reference population locations. We compared the methods using mean absolute error scaled by the true population size. Additional results are in the Online Supplement.

The median improvement in median absolute error for the latent surface model with four fixed reference groups was about 30% compared to the Zheng et al. (2006), as demonstrated in Figure 5. Thus, incorporating both the amount and the structure of the overdispersion in the network leads to better population size estimates. Additionally, the choice of how to fix group centers on the latent surface impacts the model's ability to account for the dependence relationships in the network. Fixing the centers of the four reference populations in the configuration presented in Figure 4, provides both more stable and more accurate results. These four reference groups effectively span the population and are non-overlapping, providing a high degree of structure to the space. These groups are also have high overlap with some of the remaining groups, encouraging a diverse representation of the latent structure.

4.3 Selecting the dimension of the latent surface

As with other latent geometry models, the model requires the dimension of the latent surface to be fixed in advance. Computing the posterior model probability involves the integral across all parameters in the model, which is challenging in latent geometry models (Handcock

et al., 2007; Gormley and Murphy, 2010). Further, results based on output of the MCMC likely suffer from pseudo-bias, as discussed in Lenk (2009).

In some cases it may be advantageous to predict the strength of a relationship between an individual and a certain group of alters. For example, in a link tracing sampling design such as Respondent Driven Sampling (see for example Heckathorn (1997)) there is concern the sampling process will reach a "local mode" of homogenous, highly connected individuals. Since ARD provide information about broad connectivity patterns, predictions about the associations between individuals and a given group of alters could be used to detect the position of a respondent in the network, altering researchers of a possible local mode.

We evaluate the predictive performance of our model using out-of-sample prediction on the McCarty et al. (2001) data. For each simulation, we randomly select twenty percent of the observations in the McCarty et al. (2001) dataset as our test set. Since the populations corresponding to names are used to orient the latent geometry, we do not hold out observations from these categories. We fit the algorithm described in the previous section and compute posterior predictive estimates of the missing values. We then computed the mean absolute difference between the estimated posterior medians and the held-out values. Complete results for predictive performance are displayed in the Online Supplement. Fewer latent dimensions, in general, produced a lower Mean Absolute Error (MAE) with predictive performance declining as the dimension of the latent hypersphere increased above around five. We also found that the additional structure captured by the latent surface framework leads to improved predictive performance over the Zheng et al. (2006) model, despite the additional parameters required by the latent surface model.

5 Discussion and conclusion

This paper presents a latent surface interpretation of overdispersion using ARD. We show that, through the choice of model and latent geometry, we can produce an interpretable rep-

resentation of latent network structure in survey data. Sampling from the von-Mises Fisher distribution becomes challenging in higher dimensions, though we believe this issue will have minimal impact since latent surfaces are difficult to visualize in with higher dimensions.

In conceptualizing the mapping from the full network to ARD we make assumptions about respondents' abilities to recall their network. First, we assume accurate recall from respondents' complete networks. This assumption is typically not valid for moderate to large subpopulations, though some statistical models have been proposed for similar situations (McCormick and Zheng, 2007). We also assume that the respondent has accurate information about the group membership of each of their alters. This issue, known in sociology literature as transmission errors, is more common with some groups than others (acquaintances of a diabetic may not know the person's status, for example). In some cases it is possible to select subpopulation to minimize transmission errors, yet this remains an open problem in cases where groups of interest are prone to transmission errors. Advances in data collection and survey design (e.g. Salganik et al. (2011)) offer possibilities.

Recent work demonstrates that features of network structure, such as homophily (tendency for actors to form ties with similar others), are distinguishable after aggregation (McCormick et al., 2010; McCormick and Zheng, 2012). These methods estimate network features and require detailed information about some or all subpopulations be known. The latent surface model, in contrast, provides an overview of the network dependence structure and does not require detailed subpopulation information. Comparing the latent surface model with the previous methods for ARD such as McCormick et al. (2010) or McCormick and Zheng (2012) is similar to comparing regression with principal components analysis (PCA). Regression requires model selection but gives coefficients associated with specific predictors. PCA, in contrast, requires less thought about feature selection but also produces a representation that is difficult to interpret in terms of a specific feature.

The von-Mises Fisher distribution is computationally appealing, though it imposes symmetry and unimodality. Mixtures of von-Mises Fisher distributions would provide a more

flexible representation of latent features and could provide additional insights into relationship between populations (see Mooney et al. (2003), for example). Model-based clustering across respondents could also provide information about which individuals are most likely to interact with certain groups. More importantly, it would also produce interaction profiles which reveal general patterns in link formation. A statistics literature exists for model-based clustering in spherical data (see Doret-Bernadet and Wicker (2007), for example).

References

Barber, J. J. and Gelfand, A. E. (2007). Hierarchical spatial modeling for estimation of population size. *Environmental and Ecological Statistics*, 14(3):193–205.

Braun, M. and Bonfrer, A. (2011). Scalable Inference of Customer Similarities from Interactions Data using Dirichlet Processes. *Marketing Science*, 30:513–531.

Dhillon, I. S. and Sra, S. (2003). Modeling data using directional distributions. Technical report, Technical Report TR-03-06, Department of Computer Sciences, The University of Texas at Austin.

DiPrete, T. A., Gelman, A., McCormick, T., Teitler, J., and Zheng, T. (2011). Segregation in Social Networks Based on Acquaintanceship and Trust. *The American Journal of Sociology*, 116:1234–1283.

Doret-Bernadet, J.-l. and Wicker, N. (2007). Model-based clustering on the unit sphere with an illustration using gene expression profiles. *Biostatistics*, 9:66–80.

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

Gormley, C. and Murphy, T. (2010). A mixture of experts latent position cluster model for social network data. *Statistical Methodology*, 7:385–405.

Guttorp, P. and Lockhart, R. A. (1988). Finding the location of a signal: a Bayesian analysis. *Journal of the American Statistical Association*, 83(402):322–330.

Handcock, M., Raftery, A. E., and Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society Series A*, 170:301–354.

Heckathorn, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44(2):174–199.

Hoff, P. (2009). Simulation of the Matrix Bingham–von Mises–Fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, 18:438–456.

Hoff, P. D. (2005). Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association*, 100:286–295.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97:1090–1098.

Hornik, K. and Grün, B. (2013). On conjugate families and Jeffreys priors for von Mises-Fisher distributions. *Journal of statistical planning and inference*, 143(5):992–999.

Killworth, P. D., Johnsen, E. C., Bernard, H. R., Shelley, G. A., and McCarty, C. (1990). Estimating the size of personal networks. *Social Networks*, 12:289–312.

Klovdahl, A. S., Potterat, J., Woodhouse, D. E., Muth, J. B., Muth, S. Q., and Darrow, W. (1994). Social networks and infectious disease: The colorado springs study. *Social Science and Medicine*, 38:79–88.

Lenk, P. (2009). Simulation pseudo-bias correction to the harmonic mean estimator of integrated likelihoods. *Journal of Computational and Graphical Statistics*, 18(4).

Mardia, K. V. (1972). *Statistics of directional data*. Probability and Mathematical Statistics. Academic Press, London-New York, London, 1 edition.

Mardia, K. V. and El-Atoum, S. A. M. (1976). Bayesian inference for the von Mises-Fisher distribution. *Biometrika*, 63(1):203–206.

Mardia, K. V. and Jupp, P. E. (2000). *Directional statistics*. Wiley Series in Probability and Statistics. John Wiley and Sons Ltd., Chichester, 1 edition.

McCarty, C., Killworth, P. D., Bernard, H. R., Johnsen, E. C., and Shelley, G. A. (2001). Comparing two methods for estimating network size. *Human Organization*, 60:28–39.

McCormick, T. H., Salganik, M. J., and Zheng, T. (2010). How many people do you know?: Efficiently estimating personal network size. *Journal of the American Statistical Association*, 105:59–70.

McCormick, T. H. and Zheng, T. (2007). Adjusting for recall bias in "How many X's do you know?" surveys. In *Proceedings of the Joint Statistical Meetings*.

McCormick, T. H. and Zheng, T. (2012). Latent demographic profile estimation in hard-to-reach groups. *The Annals of Applied Statistics*, 6(4):1795–1813.

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: homophily in social networks. *Annual Review of Sociology*, 27:415–444.

Mooney, J. A., Helms, P. J., and Jolliffe, I. T. (2003). Fitting mixtures of von Mises distributions: a case study involving sudden infant death syndrome. *Computational Statistics & Data Analysis*, 41:505–513.

Morris, M. (1993). Epidemiology and social networks: Modeling structured diffusion. *Sociological Methods and Research*, 22(1):99–126.

Morris, M. (2004). Network dynamism: History and lessons of the Colorado Springs Study. In Morris, M., editor, *Network Epidemiology: A Handbook for Survey Design and Data Collection*, pages 87–114. Oxford University Press, Oxford.

Rothenberg, R., Woodhouse, D., Potterat, J., Muth, S., Darrow, W., and Klovdahl, A. (1995). Social networks in disease transmission: the Colorado Springs Study. In Needle, R. H., Coyle, S., Genser, S., and Trotter, R., editors, *Social Networks, Drug Abuse, and HIV Transmission*, volume 151, pages 3–19. National Institute on Drug Abuse.

Salganik, M. J., Mello, M. B., Adbo, A. H., Bertoni, N., Fatzio, D., and Bastos, F. I. (2011). The game of contacts: Estimating the social visibility of groups. *Social Networks*, 33:70–78.

Shelley, G., Bernard, H., Killworth, P., Johnsen, E., and McCarty, C. (1995). Who knows you HIV status? What HIV+ patients and their network members know about each other. *Social Networks*, 17:189–217.

Smyth, G. K. (1998). Polynomial approximation. In Armitage, P. and Colton, T., editors, *Encyclopedia of Biostatistics*. Wiley.

UNAIDS (2003). *Estimating the Size of Popualtions at Risk for HIV*. Number 03.36E in Issues and Methodology. UNAIDS, Geneva.

USDHHS (2013). National survey of adoptive parents.

Wood, A. T. A. (1994). Simulation of the von mises fisher distribution. *Communications in statistics-simulation and computation*, 23:157–64.

Woodhouse, D., Rothenberg, R., Potterat, J., Darrow, W., Muth, S., Klovdahl, A., Zimmerman, H., Rogers, H., Maldonado, T., and et al., J. M. (1994). Mapping a social network of heterosexuals at high risk for HIV infection. *AIDS*, 8:1331–1336.

Zheng, T., Salganik, M. J., and Gelman, A. (2006). How many people do you know in prison?: Using overdispersion in count data to estimate social structure. *Journal of the American Statistical Association*, 101:409–423.