# A latent space representation of overdispersed relative propensity in "How many X's do you know?" data

Tyler H. McCormick* Tian Zheng†

**Abstract**

We present a novel latent space representation of the relative propensity for a respondent to form ties with members of a particular social group, a quantity related to overdispersion. In many applications collecting complete network data is financially or practically infeasible. Instead, we use data where respondents are asked for the number of ties they have with members of various subpopulations or "How many X's do you know?" data. We connect this data with recent work using models which represent dependence in a fully observed network through distance in an unobservable "social space," known as latent space models (Hoff et al., 2002). This yields a latent space representation of overdispersion, further elucidates how these data measure social structure indirectly, and suggests a latent space model for such data.

**Key Words:** How many X's do you know? data, latent space models, overdispersion, social networks

## 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 increasingly common in the social and behavioral sciences and typically contain higher order dependence structure. This issue has given rise to a number of statistical models, with one recent attempt being the family of latent space models first applied to networks in Hoff et al. (2002). The latent space model assumes that the actors in the network form ties independently given their (latent) position in some unobservable "social space."

In this paper we derive a latent space interpretation of overdispersion in "How many X's do you know?" data. 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 subopulation. We measure overdispersion using data collected through standard surveys, known as "How many X's do you know?" data. Here, $X$, represents a subpopulation of interest. These subpopulations often include first names (2006 GSS, McCarty et al. (2001)). First names are particularly useful in learning about network structure since many aggregate features of alters with a given name are available from the Census Bureau and Social Security Administration. Other potential $X's$ may be of interest in their own right. McCarty et al. (2001) also asks about individuals who are HIV positive and the UNDP currently sponsors several projects that ask about behaviors they deem risk factors for contracting HIV/AIDS. Both McCarty et al. (2001) and the 2006 GSS module also asked about particular occupations and life situations. These data are often used to learn about populations which are difficult to reach using standard surveys (see Killworth et al.

*Department of Statistics, Columbia University, 1255 Amsterdam Ave, New York, NY 10025. This author is supported by a Google Ph.D. Fellowship in Statistics.

†Department of Statistics, Columbia University, 1255 Amsterdam Ave, New York, NY 10025

(1998) for example) and have recently also been used to learn about other social science phenomena (DiPrete et al., 2010).

Defining "know" (or another relationship such as trust) defines the network of interest. Given this network, "How many X's do you know?" data are a type of network sample. If respondents could recall perfectly from their network and had full knowledge of all of the group memberships of all alters, then these data would be "equivalent" to asking a respondent if they know each member of a particular group of alters. If every Michael in the US population were standing in a room, for example, we could imagine asking the respondent if he/she has a tie with each person in the room. 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.

Recent work with this data demonstrates that features of network structure, such as homophily (the tendency for actors to form relationships with similar others), are distinguishable even after the aggregation described above (McCormick et al., 2010). Along with developing a latent space representation for overdispersion, we also extend this literature by providing a specific pathway from a common complete network latent space model to a model for "How many X's do you know?" data.

The remaining sections are organized as follows. In Section 2 we describe a latent space model for cases when the entire network are observed. Next, in Section 3 we derive the mathematical relationship between the complete network model presented in Section 2 and a latent space model for "How many X's do you know?" data and present a latent space interpretation of overdispersion. Section 4 gives a discussion of future directions.

## 2. A latent space model for complete graphs

Consider two actors $i$ and $j$ whose relationship is described by the sociomatrix $\Delta$ where $\delta_{ij} = 1$ if there is a link between $i$ and $j$ and 0 otherwise. Let the gregariousness be distributed $g_i \sim \mathrm{F}(\mu_g, \sigma_g)$ for an actor $i$ and for any member, $j$, of subpopulation $k$, $g_{j\in k} \sim \mathrm{F}(\mu_{g_k}, \sigma_{g_k})$. Group-dependent gregariousness distributions accommodates variability in overall tie frequency associated with some subpopulations. Politicians and members of the clergy typically have above-average degree, for example (McCarty et al., 2001). Many hard-to-count subpopulations may display below-average connectivity. Say there are $N_k$ members of subpopulation $k$ and $N$ members of the population. Let $\mathcal{S}^p$ be the $p$ dimensional hypersphere. $\mathbf{z}_i$ and $\mathbf{z}_j$ are the latent position vectors of $i, j$ on $\mathcal{S}^{p+1}$, corresponding to a $p$ dimensional latent space on the $p+1$ dimensional hypersphere.

Consider now a linear latent space model similar to the one presented by Hoff (2005). That is:

$$\begin{aligned} \theta_{ij} &= g_i + g_j + \eta \mathbf{z}_i' \mathbf{z}_j \\ E(\delta_{ij}|\theta_{ij}) &= \mathrm{h}(\theta_{ij}) \end{aligned}$$

where $h(\cdot)$ is the link function. The self-closure property of the hypersphere facilitates putting uniform distributions on $\mathbf{z}_i$. Under this assumption and conditional on a tie between two actors, the distribution of $\mathbf{z}_j$

$$\begin{aligned} P(\mathbf{z}_j|\delta_{ij} = 1, \mathbf{z}_i) &= \frac{P(\delta_{ij} = 1|\mathbf{z}_i, \mathbf{z}_j)P(\mathbf{z}_j)}{P(\delta_{ij} = 1|\mathbf{z}_i)} \\ &= C_{p+1}(\eta) \exp(\eta \mathbf{z}_i' \mathbf{z}_j), \end{aligned}$$

which is $\mathcal{M}(\mathbf{z}_i, \eta)$. $\mathcal{M}(\mathbf{z}_i, \eta)$ corresponds to a von-Mises Fisher distribution with mean $\mathbf{z}_i$ and concentration $\eta$. Thus, $\eta$ models the overall level of homophily in the population with individuals distributed uniformly. For members of specific subpopulations, $\mathbf{z}_{j\in k} \sim \mathcal{M}(\boldsymbol{\upsilon}_k, \eta_k)$ for members of subpopulation $k$, or a von-Mises Fisher distribution with center $\upsilon_k$ and concentration $\eta_k$. Higher values of $\eta_k$ correspond to distribution with more mass concentrated around $\upsilon_k$.

Specifically, we consider the log-linear model such that

$$\mathrm{E}(\delta_{ij}|g_i, g_j, \eta, \mathbf{z}_i, \mathbf{z}_j) = \exp(g_i + g_j + \eta \mathbf{z}_i' \mathbf{z}_j). \tag{1}$$

## 3. Latent space representation of overdispersion

As described in Section 1 a simple conceptualization of "How many X's do you know?" data involves asking a respondents if they know every member of a set of subpopulations then reporting only the aggregate number known in that subpopulation. We begin by computing the expected number of individuals known by a respondent in a given subpopulation. Next, we derive representations of two network features: degree and fractional subpopulation proportion. We then relate these quantities through the model for overdispersion presented in Zheng et al. (2006). From this relationship we relate features of the latent space to overdispersion.

Consider "How many X's do you know?" data $y_{ik}$ where, for a subpopulation $k$, $y_{ik}$ relates to the full network model in Section 2 as $y_{ik} = \sum_{j\in k} \delta_{ij}$. Conditioning on $\mathbf{z}_i$ and $\mathbf{z}_{j\in k}$, $\delta_{ij}$ are independent Bernoulli trials each with a small probability of being 1. When $N_k$ is large, which is usually the case, $y_{ik}$ follows the Poisson distribution with rate $\lambda_{ik} = \sum_{j\in k} P(\delta_{ij} = 1) = N_k P(\delta_{ij} = 1)$. As discussed in Zheng et al. (2006), the super-Poisson variation present in the $y_{ik}$'s, or overdispersion, is defined by variation in individuals' relative propensity to form ties with members of certain subpopulations that cannot be explained by degree variation and group sizes. Therefore, we focus on $\lambda_{ik}$ and examine its relationship with degree and fractional subpopulation size. Beginning by defining $\lambda_{ik}$ as the expectation of our observed data,

$$\begin{aligned}
\mathrm{E}(y_{ik}) &\triangleq \lambda_{ik} \\
&= \mathrm{E}\sum_{j\in k} \delta_{ij} = \sum_{j\in k} \mathrm{E}(\delta_{ij}) \\
&\approx N_k \int_{\mathbf{z}_{j\in k}} \exp(g_i + g_j + \eta \mathbf{z}_i' \mathbf{z}_j) P(\mathbf{z}_j) P(g_j) d\mathbf{z}_j dg_j \\
&= N_k \exp(g_i) \mathrm{E}_k(\exp(g_j)) \int_{\mathbf{z}_{j\in k}} \exp(\eta \mathbf{z}_i' \mathbf{z}_j) C_{p+1}(\eta_k) \exp(\eta_k \boldsymbol{\upsilon}_k' \mathbf{z}_j) d\mathbf{z}_j \\
&= N_k \exp(g_i) \mathrm{E}_k(\exp(g_j)) C_{p+1}(\eta_k) \int_{\mathbf{z}_{j\in k}} \exp\left((\eta \mathbf{z}_i + \eta_k \boldsymbol{\upsilon}_k)' \mathbf{z}_j\right) d\mathbf{z}_j \\
&= N_k \exp(g_i) \mathrm{E}_k(\exp(g_j)) C_{p+1}(\eta_k) \int_{\mathbf{z}_{j\in k}} \exp\left(\frac{||\eta \mathbf{z}_i + \eta_k \boldsymbol{\upsilon}_k||}{||\eta \mathbf{z}_i + \eta_k \boldsymbol{\upsilon}_k||} (\eta \mathbf{z}_i + \eta_k \boldsymbol{\upsilon}_k)' \mathbf{z}_j\right) d\mathbf{z}_j
\end{aligned}$$

The integral now contains the kernel of $\mathcal{M}\left(||\eta \mathbf{z}_i + \eta_k \boldsymbol{\upsilon}_k||, \frac{\eta \mathbf{z}_i + \eta_k \boldsymbol{\upsilon}_k}{||\eta \mathbf{z}_i + \eta_k \boldsymbol{\upsilon}_k||}\right)$, which yields

$$\lambda_{ik} = N_k \exp(g_i) \mathrm{E}_k(\exp(g_j)) \left(\frac{C_{p+1}(\eta_k)}{C_{p+1}(||\eta \mathbf{z}_i + \eta_k \boldsymbol{\upsilon}_k||)}\right)$$

Moving now to the expected respondent degree:

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

Since $\exp(\eta \mathbf{z}_i' \mathbf{z}_j)$ is the kernel of $\mathcal{M}(\eta, \mathbf{z}_j)$ we have

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

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

$$d_i = N \exp(g_i) \mathrm{E}(\exp(g_j)) \left( \frac{C_{p+1}(0)}{C_{p+1}(\eta)} \right). \tag{2}$$

Now moving to the fractional subpopulation size:

$$\begin{aligned} \beta_k &\triangleq \frac{\sum_i \sum_{j \in k} \delta_{ij}}{\sum_{ij} \delta_{ij}} \\ \text{where} & \\ \sum_{ij} \delta_{ij} &= N \mathrm{E}(\sum_j \delta_{ij}) = N \mathrm{E}(d_i) \\ &= N \int_i N \exp(g_i) \mathrm{E}(\exp(g_i)) \frac{C_{p+1}(0)}{C_{p+1}(\eta)} p(g_i) dg_i \\ &= N^2 (\mathrm{E}(\exp(g_i))^2 \frac{C_{p+1}(0)}{C_{p+1}(\eta)} \end{aligned}$$

and

$$\begin{aligned}
\sum_i \sum_{j\in k} \delta_{ij} &= \sum_i N_k \mathrm{E}(\delta_{ij}|j \in k) \\
&= \sum_i N_k \int_{j\in k} \exp(g_i + g_j + \eta \mathbf{z}_i'\mathbf{z}_j) p(g_j) p(\mathbf{z}_j) d\mathbf{z}_j dg_j \\
&= \sum_i N_k C_{p+1}(\eta_k) \exp(g_i) \mathrm{E}_k(\exp(g_j)) \int_{j\in k} \exp(\eta_n \boldsymbol{\upsilon}_k'\mathbf{z}_j + \eta \mathbf{z}_i'\mathbf{z}_j) d\mathbf{z}_j \\
&\quad \text{taking the expectation over } i \\
&= N\mathrm{E}\left( N_k C_{p+1}(\eta_k) \exp(g_i) \mathrm{E}_k(\exp(g_j)) \int_{j\in k} \exp(\eta_n \boldsymbol{\upsilon}_k'\mathbf{z}_j + \eta \mathbf{z}_i'\mathbf{z}_j) d\mathbf{z}_j \right) \\
&= N N_k C_{p+1}(\eta_k) \mathrm{E}(\exp(g_i)) \mathrm{E}_k(\exp(g_j)) C_{p+1}(0) \int_i \int_{j\in k} \exp(\eta_k \boldsymbol{\upsilon}_k'\mathbf{z}_j + \eta \mathbf{z}_i'\mathbf{z}_j) d\mathbf{z}_j d\mathbf{z}_i \\
&\quad \text{exchanging order of integration} \\
&= N N_k C_{p+1}(\eta_k) \mathrm{E}(\exp(g_i)) \mathrm{E}_k(\exp(g_j)) C_{p+1}(0) \int_{j\in k} \exp(\eta_k \boldsymbol{\upsilon}_k'\mathbf{z}_j) \left( \int_i \exp(\eta \mathbf{z}_i'\mathbf{z}_j) d\mathbf{z}_i \right) d\mathbf{z}_j \\
&\quad \text{symmetry of inner product makes } \exp(\eta \mathbf{z}_i'\mathbf{z}_j) \text{the kernel of } \mathcal{M}(\eta, \mathbf{z}_j) \\
&= N N_k C_{p+1}(\eta_k) \mathrm{E}(\exp(g_i)) \mathrm{E}_k(\exp(g_j)) \frac{C_{p+1}(0)}{C_{p+1}(\eta)} \int_{j\in k} \exp(\eta_k \boldsymbol{\upsilon}_k'\mathbf{z}_j) d\mathbf{z}_j \\
&\quad \text{where } \exp(\eta_k \boldsymbol{\upsilon}_k'\mathbf{z}_j) \text{is kernel of } \mathcal{M}(\eta_k, \boldsymbol{\upsilon}_k) \\
&= N N_k C_{p+1}(\eta_k) \mathrm{E}(\exp(g_i)) \mathrm{E}_k(\exp(g_j)) \frac{C_{p+1}(0)}{C_{p+1}(\eta) C_{p+1}(\eta_k)} \\
&= N N_k \mathrm{E}(\exp(g_i)) \mathrm{E}_k(\exp(g_j)) \frac{C_{p+1}(0)}{C_{p+1}(\eta)}.
\end{aligned}$$

Thus, after combining the two pieces,

$$\beta_k = \left( \frac{N_k}{N} \right) \left( \frac{\mathrm{E}_k(\exp(g_j))}{\mathrm{E}(\exp(g_i))} \right)$$

We subsitute $\frac{C_{p+1}(0)}{C_{p+1}(\eta)} d_i \beta_k = N_k \exp(g_i) \mathrm{E}_k(\exp(g_j))$ we have

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

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

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

which corresponds to the Zheng et al. (2006) overdispersed model with $\lambda_{ik} = d_i \beta_k \gamma_{ik}$ where $\gamma_{ik}$ controls the relative propensity for $i$ to form ties with group $k$ and is given by

$$\gamma_{ik} = \frac{C_{p+1}(\eta) C_{p+1}(\eta_k)}{C_{p+1}(0) C_{p+1}(\sqrt{\eta^2 + \eta_k^2 + 2\eta\eta_k \cos(\theta_{(\mathbf{z}_i, \boldsymbol{\upsilon}_k)})})}. \tag{3}$$

In the latent space model for the full network (see (1)), the latent space component $(\eta \mathbf{z}_i'\mathbf{z}_j)$ increases the propensity for individuals who are more similar in the unobserved social space to interact, which corresponds to a form of non-random mixing.

If the $\eta$ parameter were zero, however, we would be left with a model that accounts for varying gregariousness across actors but assumes random mixing across all other attributes. Setting $\eta = 0$ in (3), we see have that $\gamma_{ik} = 1$ and the model simplifies to the "null model" for random mixing presented in Zheng et al. (2006).

Rather than estimating $\gamma_{ik}$ directly, Zheng et al. (2006) assign a Gamma prior distribution to $\gamma_{ik}$ with a mean of 1 and shape parameter $1/(\omega_k - 1)$. The $\gamma$'s can then be integrated out to yield a Negative Binomial distribution with overdispersion parameter $\omega_k$. In the latent space representation, taking the expectation of $\lambda_{ik}$ and rearranging the resulting expression yields that $\gamma_{ik}$ has expectation 1. The variance of $\gamma_{ik}$ (and therefore overdispersion) increases monotonically as the concentration of the subpopulation $\eta_k$ increases relative to the general level of the population, $\eta$. This result can be verified through simulation (not shown).

## 4. Discussion

We present a latent space interpretation of overdispersion using "How many X's do you know?" data. We begin with a latent space model for the full network, then aggregate across various subpopulations of interest. We then relate this aggregation to overdispersion models for this type of data presented in Zheng et al. (2006).

In conceptualizing the mapping from the full network to "How many X's do you know?" data we make assumptions about respondents' abilities to recall their network. First, we assume that respondents recall accurately from their complete network. 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 subpopulations 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 subpopulations of interest are prone to transmission errors.

## References


DiPrete, T. A., Gelman, A., McCormick, T. H., Teitler, J., and Zheng, T. (2010). Segregation in social networks based on acquaintanceship and trust. *American Journal of Sociology*, To appear.

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.

Killworth, P. D., Johnsen, E. C., McCarty, C., Shelley, G. A., and Bernard, H. R. (1998). A social network approach to estimating seroprevalence in the United States. *Social Networks*, 20:23–50.

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 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*.

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.