

Latent Structure Models for Social Networks using Aggregated Relational Data

Tyler H. McCormick
tyler@stat.columbia.edu

Tian Zheng
tzheng@stat.columbia.edu

Department of Statistics, Columbia University

Abstract

We propose a model where the propensity for an individual to know members of a given alter group (people named Michael, for example) is independent given the positions of the individual and the group in a latent “social space.” Using this framework, we derive evidence of social structure in personal acquaintance networks, estimate homogeneity of groups, and estimate individual and population gregariousness. We apply our method to data from McCarty et al. (2001).

Latent space models for networks

- Presence/absence of a tie between two members of the network, i and j , is independent of the other ties in the network given the positions of i and j in a latent “social space.”
- “Social Space refers to a space of unobserved latent characteristics that represent potential transitive tendencies in network relations.” (Hoff et al. 2002)

Aggregated Relational Data

- Complete network data are often difficult to collect, especially in the social sciences.
- We propose a latent space model for ties observed indirectly using questions of the form:

How many X’s do you know?

- Instead of learning about an respondent’s relationship with one other member of the network we learn about his/her relationship with a group.
- We use data from McCarty et al. (2001). There are 1375 respondents and twelve names. Demographic information about names is available from U.S. Social Security Administration.
- Additional subpopulations include: HIV/AIDS, homeless, prison, Jaycees, adopt a child, dialysis, died in auto accident in the past year, commercial pilots.

Latent Geometry

We propose using a p dimensional latent space mapped to the $p + 1$ dimensional unit sphere, \mathcal{S}^{p+1} . We define distance as the inner product on \mathcal{S}^{p+1} (great-circle distance, geodesic distance). The geometry of the sphere facilitates uniform priors on respondent latent positions and the inner product distance has an interpretation similar to a random effect term.

Latent space models for ARD

Define y_{ik} as the number of people respondent i knows in subpopulation k . We model $y_{ik} \mid \lambda_{ik} \sim \text{Poisson}(\lambda_{ik})$

$$\lambda_{ik} = d_i b_k \mathbb{E}_{z_{j \in k}} (\exp(\eta z_i' z_{j \in k}))$$

d_i and b_k , the degree and fractional subpopulation size, respectively, have Gaussian priors on the log scale. η has a Gamma prior and modulates the intensity of latent influence. We model latent positions on \mathcal{S}^{p+1} as:

$$\begin{aligned} z_i \mid \mu_z, \eta_z &\sim \mathcal{M}(\mu_z, 0) \\ z_{j \in k} \mid \mu_k, \eta_k &\sim \mathcal{M}(\mu_k, \eta_k) \end{aligned}$$

where \mathcal{M} denotes the von Mises-Fisher distribution across points on \mathcal{S}^{p+1} . Since the receivers, $j \in k$, are not observed directly, we cannot model their latent positions explicitly. Instead, we estimate the latent distribution of members of group k .

Aggregation and latent inference

- Aggregated Relational Data are a specific type of sample from a social network, “equivalent” to asking respondents if they know each member of a group.
- We don’t observe receivers, $j \in k$, so there’s no way to determine their position in the latent space, instead we make inferences on the expected distance between respondents and the center of the subpopulation.
- Implications of aggregation are methodologically and sociologically relevant.
- Consider two actors i and j whose relationship is described by the sociomatrix Δ where $\delta_{ij} = 1$ if there is a link between i and j and 0 otherwise.
- Assume latent space model in the generalized linear model framework similar to Hoff (2005). Specifically, we consider the log-linear model such that

$$\mathbb{E}(\delta_{ij} \mid g_i, g_j, \eta, z_i, z_j) = \exp(g_i + g_j + \eta z_i' z_j).$$

- Gregariousness has distribution $g_i \sim F(\mu_g, \sigma_g)$ for the overall population and $g_{j \in k} \sim F(\mu_{g_k}, \sigma_{g_k})$ for members of subpopulation k .
- Conditioning on z_i and $z_{j \in k}$, δ_{ij} are independent Bernoulli trials, each with a small success probability.
- These probabilities vary amongst alters $j \in k$ depending on the spread of the distribution of $z_{j \in k}$.
- Since receivers, $j \in k$, are not observed, we instead use the expectation, which implies that $y_{ik} = \sum_{j \in k} \delta_{ij} = \sum_k \mathbf{1}(i \text{ knows } j)$ approximately follows a Poisson distribution.
- Then compute the form of the Poisson rate,

$$\begin{aligned} \mathbb{E}(y_{ik}) &\triangleq \lambda_{ik} \\ &= \mathbb{E} \sum_{j \in k} \delta_{ij} \approx N_k \int_{z_{j \in k}} \exp(g_i + g_j + \eta z_i' z_j) P(z_j) P(g_j) dz_j dg_j. \end{aligned}$$

References

- [1] Hoff, P. D. (2005), “Bilinear Mixed-Effects Models for Dyadic Data,” *Journal of the American Statistical Association*, 100, 286-295.
- [2] 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.
- [3] McCarty, C., Killworth, P. D., Bernard, H. R., Johnsen, E., and Shelley, G. A. (2001), “Comparing Two Methods for Estimating Network Size,” *Human Organization*, 60, 28-39.
- [4] 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.