Caste, access to credit and social engagement: a network theoretic approach

Advait Rajagopal

December 18, 2017

Overview

- Analyzing networks has the potential to provide insight into human behavior and interactions. Gaining an understanding of the structures of networks, the central agents in networks and the density of relationships can shed light on how information spreads in a village and ultimately how people behave (Jackson 2014). Cultural factors and institutions in rural development are intertwined (Arora and Sanditov 2015) and factors like caste have the potential to impact social relations.
- The goal of this project is to study the borrowing behavior in villages in Karnataka. I use borrowing as a proxy for "access to credit". I want to study the relationship between borrowing and social engagement in villages while explicitly accounting for caste level variation. By obtaining caste level coefficients for the relation between social engagement and borrowing, we can understand whether some castes have a higher coefficient associated with a given level of social engagement suggesting that this predicts higher access to credit.
- In order to obtain an overall regression coefficient as well as caste level coefficients simultaneously, I use Bayesian hierarchical model with partial pooling across castes (see Gelman et al. (2013), ch. 5 for a clear explanation of Bayesian hierarchical models and partial pooling). The dependent variable is the degree (number of connections each node has) of each node in the borrowing network and the predictor is the degree of each node in the social engagement network.
  - Relevant questions asked to the respondents: "whom did you borrow money from?", "whom do you socially engage with?"
- Important reasons for using a Bayesian hierarchical model with partial pooling include
  - In enables me to use the hierarchical modeling structure to get both overall regression coefficients and caste level coefficients at the same time
  - Once nodes without covariate information are dropped, the sample size within group (i.e within caste) can become very small. Using informative prior information can help with inference in cases where there are very small samples in a subgroup.
  - Partially pooling across castes allows me to model my belief that nodes belonging to different castes are neither exactly the same, nor completely different.

The dataset

- The network dataset is compiled by the MIT based "Abdul Latif Jameel Poverty Action Lab". It is a network dataset that contains relational data about borrowing, lending, social activities and other behavioral patterns. It also has demographic data that includes information about occupation, religion, caste, education, migration patterns and a lot more. There is data for 77 villages, containing household level information and individual level information. The data is available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/21538

Limitations

- Many of the questions that form the adjacency matrices for the networks in this dataset are one dimensional and do not indicate the direction of the relationship between the two nodes. For example,

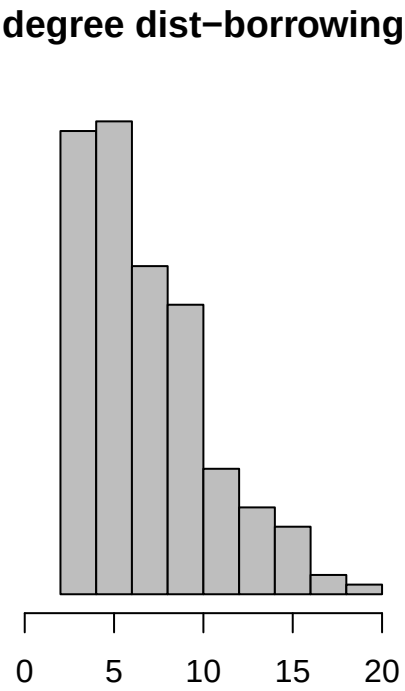
consider the question, "whom did you borrow money from?", we don't know who the borrower is and who the lender is, and matrices are considered symmetric. If $y_{ij} = 1$ in the matrix then I interpret this as individual i and j have a relationship where they offer each other credit.
- Missing covariate information for most of the nodes in the network has also been problematic. This should be addressed and accounted for in further expansions of this study with appropriate weighting etc.
- The only covariate I account for is caste. This is because I want to set up the modeling structure and understand relationships before controlling for additional covariate information. This is an essential next step in the process and I hope to do this in future
- Errors in this document are all mine, please reach out to me at rajaa598@newschool.edu for more information, suggestions, and corrections.

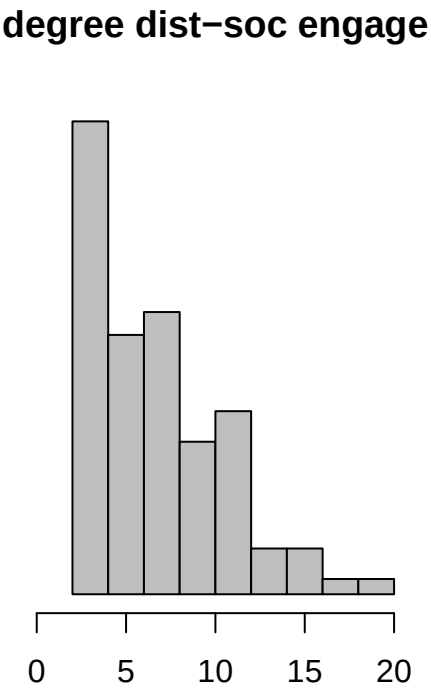
How this poster is organized

- I have arbitrarily selected village 74 (the dataset is huge and this is an experiment that I hope to scale to other villages in the future)
- First, I explore the dataset, color the nodes according to the caste attributes and set up network graphs.
- Second, I set up models to reflect some of the patterns in the dataset and my choice of probability distribution reflects the nature of the distribution of degree centrality in "borrowing" and "social engagement" networks.
  - I set up a Normal model to predict degree in "borrowing" networks. I do this because the normal distribution is easy to udnerstand and the coefficient is easy to interpret. The downside is that the support of the Normal distribution is from $-\infty$ to $+\infty$ and it is a continuous distribution. These features are not ideal for non-negative discrete data, i.e degree centrality of nodes in "borrowing" networks.
  - I set up a Poisson model to predict degree in "borrowing" networks. The Poisson is correct for modeling discrete values, but the coefficients are harder to interpret.
- Lastly, I validate my models with posterior predictive checks and interpret the results of the modeling exercise and conclude.

Exploratory data analysis

- The datasets I use are the:
  - individual covariate information
  - borrowing money in village 74 (adjacency matrix indicating who borrowed money from whom)
  - social engagement in village 74 (adjacency matrix indicating who socially engages with whom)
- In order to explore the datasets, we want to first understand how many people/nodes in the village actually have covariate information. Data collection in the field and in developing countries can often be difficult and so there is a lot of missing covariate information. Ultimately I find that there are 193 nodes with covariate information
- I then make the necessary network objects. Once I have changed the simple adjacency matrices into network objects, I am able to plot the degree distribution. Below we can see the degree distributions of these networks. The degree distribution appears to be positively skewed with few popular or high degree nodes and most nodes with a low degree.





- Now I examine the covariate information and add the "caste" information as a node level covariate to the network object. After adding the node level covariate as an attribute, I summarize the important network statitics below.

Borrowing network (dependent variable)

Borrowing network	Value
Vertices	193
Total Edges	372
Missing Edges	0
Non-missing Edges	372
Density	0.02

Social Engagement network (independent variable)

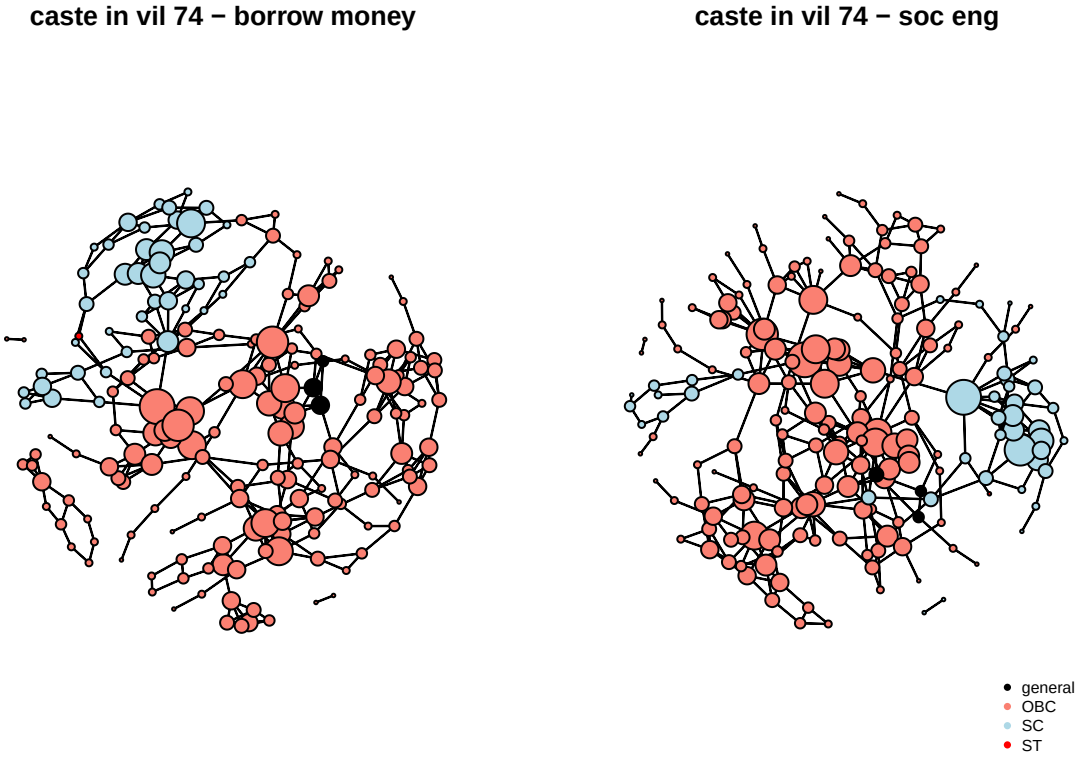
Social Engagement network	Value
Vertices	193
Total Edges	362
Missing Edges	0
Non-missing Edges	362
Density	0.02

Covariate

- Caste attribute of the nodes;

Caste attribute	Value
General	3
OBC	147
SC	42
ST	1
Total	193

- I plot the networks of borrowing and social engagement side by side with nodes colored to represent caste. The node sizes reflect the degree of each node. We observe that there are mostly OBC (other backward class) and SC (scheduled caste) nodes with very few general and just one ST (scheduled tribe) node. OBC, general/forward, SC, ST account for roughly 41%, 31%, 19% and 9% of the Indian population total. It is surprising that even though there are only 3 "general" nodes they appear to be quite popular and have decent access to credit. This becomes even more significant at the time of modeling. The one ST node is also a potential challenge to modeling as inference is difficult with smaller sample sizes.



The "general" nodes are black, "OBC" are salmon, "SC" are light blue and "ST" is red.

Hierarchical Models

- I set up a normal model and a Poisson model to capture the relationship between social engagement and access to credit.
- As mentioned in the introduction, the normal model fits a hierarchical linear regression and is useful because it has easily interpretable coefficients. Each coefficient can be seen as the effect on the outcome variable. The drawbacks of the normal distribution for modeling degrees of nodes are that, (i) the support of the normal distribution extends into the negative half of the real line and (ii) it is a continuous distribution, and we are modeling discrete degrees of nodes, which is not strictly correct.

- The Poisson model is appropriate because we model a discrete distribution for count data and given that we are modeling degree distributions this seems more correct.
- The Poisson model performs marginally better than the Normal model in its ability to predict the longer right hand tail of the degree distribution and the positively skewed nature of the degree distribution.
- I use Stan (Carpenter et al. 2017) to sample from the joint posterior of the observed and unobservable quantities. Stan uses a Hamiltonian Monte Carlo (HMC) algorithm (one from a broader class of MCMC sampling methods) (Betancourt 2017) to accomplish this. The posterior distribution is proportional to the product of the prior and likelihood.

Normal model for caste level effects

Model fitting

The full model is described in the equations below;

$$
\begin{aligned}
p(deg_{bor}|deg_{social}, \beta_0, \beta_1) &\sim \mathcal{N}(\beta_0 + \beta_{1_c}.deg_{social_c}, \sigma^2) \\
\beta_0 &\sim \mathcal{N}(0, 1) \\
\beta_{1_c} &\sim \mathcal{N}(\mu_c, \tau_c) \\
\mu_c &\sim \mathcal{N}(0, 0.5) \\
\tau_c &\sim Cauchy^{+}(0, 10) \\
\sigma &\sim Cauchy^{+}(0, 10)
\end{aligned}
$$

The first equation in this set is the likelihood function (the probability of the dependent variable conditional on the independent variable and the model parameters). The rest of the equations are priors on the parameters.

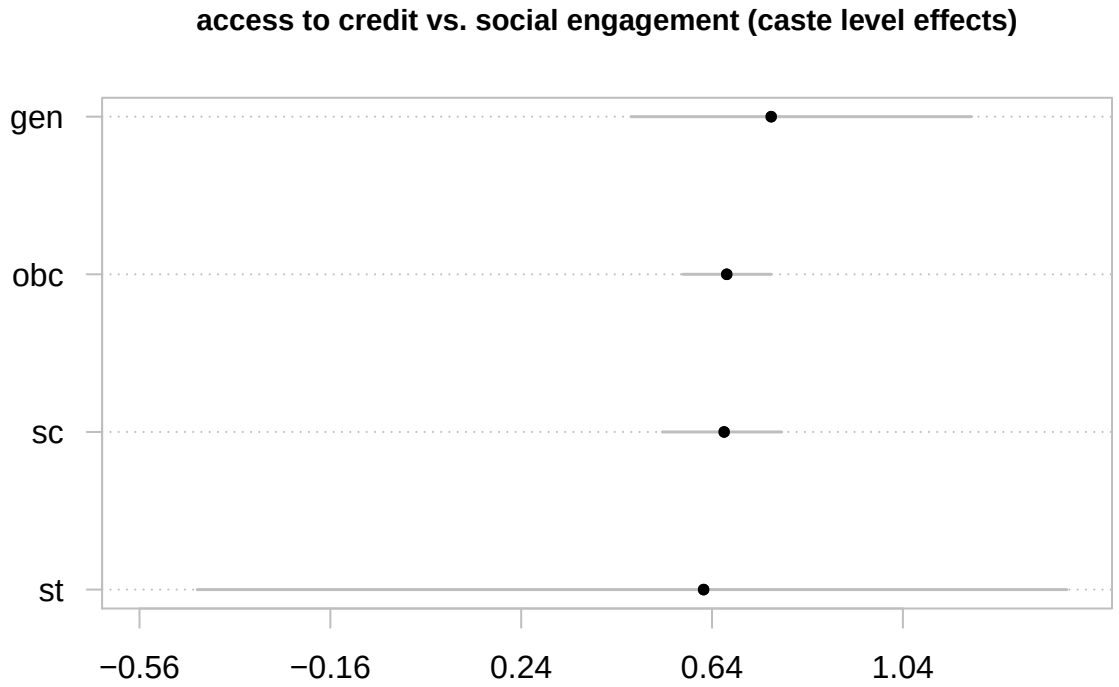
I fit a normal model where β_1 is the slope coefficient that measures the caste specific marginal effect of social engagement degree on the degree in the borrowing network. I use weakly informative (Gelman, Simpson, and Betancourt 2017) normal priors on the parameters β_0, β_1 and μ_c. Gelman et al. (2006) recommend a half-Cauchy prior on variances τ_c and σ as this distribution has most of its mass at the mean 0 but has smooth slopes and long tails placing some mass significantly away from 0 allowing variance to creep into the model, but simultaneously constraining the space the chains of the HMC have to explore.

Posterior distribution of the estimated parameters

	mean	se_mean	sd	2.5%	97.5%	n_eff	Rhat
b0	2.60	0.0230	0.400	1.900	3.40	290	1
b1[1]	0.76	0.0093	0.180	0.470	1.20	370	1
b1[2]	0.67	0.0026	0.049	0.580	0.76	350	1
b1[3]	0.67	0.0032	0.063	0.540	0.79	390	1
b1[4]	0.62	0.0210	0.430	-0.440	1.40	420	1
mu_c	0.60	0.0120	0.230	-0.042	0.91	350	1
tau_c	0.34	0.0380	0.550	0.031	1.60	210	1

Based on the above table we have obtained the posterior distribution for β_1 and there are four of these for each caste. We can see the 95% posterior interval for the parameters. We see the model converges well as the Rhat (Carpenter et al. 2017) values are very close to 1.

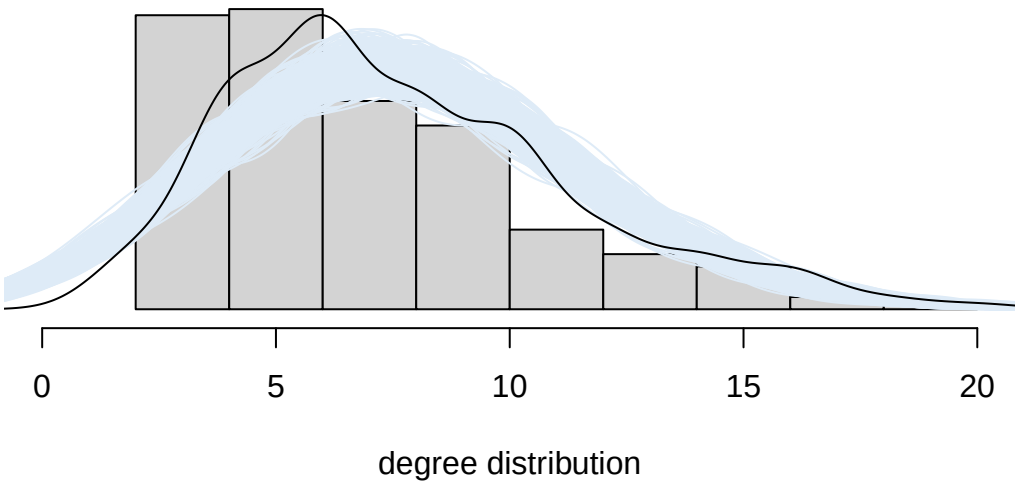
I plot the marginal posterior distributions of the parameters below.



- The posterior intervals of the β_1 parameters and there are 4 (one for each caste) corresponding to "general", "OBC", "SC" and "ST" respectively.
- We see that even with three data points the model picks up on the fact that being of the "general" category predisposes you to have a higher degree in the borrowing network. This suggests that the historically "upper" caste, despite being a minority has more access to credit. This is potentially an important result. It is worthwhile to check this with other networks that have different compositions of castes to make sure that this conclusion holds there as well.
- The "general" and "ST" coefficients having very large error bars or uncertainty intervals. This is because there are only 3 general nodes and 1 ST node.
- It is interesting that despite the small samples in these two groups we are able to obtain stable regression coefficients. We also find that the general group's coefficient is highest. The general, SC and OBC posterior distributions exclude 0, and the ST interval includes 0. This suggests that there may be an effect for the first three but not for ST. However this could also be because there is only one ST node and inherently brings in uncertainty into the estimate. The model needs to be scaled to other networks to examine the patterns.

Posterior predictive checking

I now engage in posterior predictive checking. Using my normal model I predict the degree distribution for the borrowing network and compare it to the raw data. I plot the histogram of the raw data. The light blue lines are simulated values for the dependent variable (degree of nodes in the borrowing network). The black line is the true density distribution of the dependent variable.



- We see that the normal distribution has a longer right tail, but it cannot quite capture the positively skewed degree distribution. The mean of the simulations is slightly to the right of the true distribution. This is because we have modeled the distribution of degrees as normal when it is clearly poisitively skewed and the outcome variable is non negative integral values.

Poisson model for caste level effects

Model fitting

The full model is described in the equations below;

$$
\begin{aligned}
p(deg_{bor}|deg_{social}, \beta_1) &\sim \text{Poisson}(\beta_{1_c}.deg_{social_c}) \\
\beta_{1_c} &\sim \mathcal{N}(\mu_c, \tau_c) \\
\mu_c &\sim \mathcal{N}(1, 1) \\
\tau_c &\sim Cauchy^{+}(0, 10)
\end{aligned}
$$

Like the previous set of equations, the first equation here is the likelihood function and the other equations are priors on the parameters. Here the Poisson parameter is a function of the predictor variable. This model is appropriate because the Poisson distribution is discrete and does not support the negative half of the real line. This accurately reflects the nature of the degree of a node in borrowing networks.

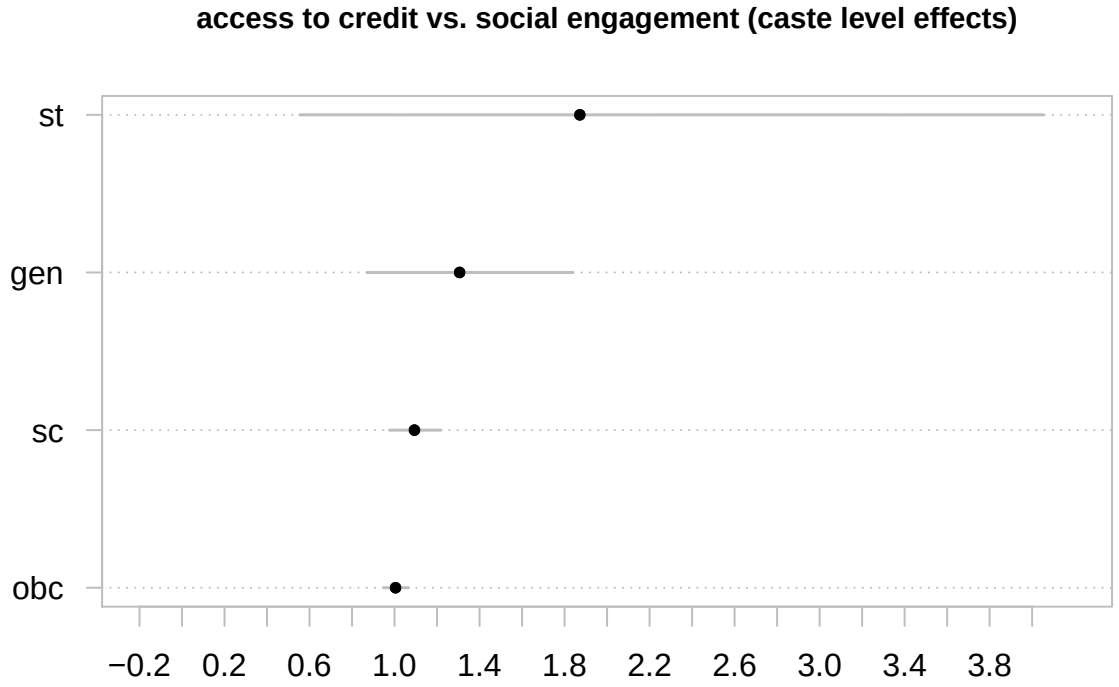
Posterior distribution of the estimated parameters

	mean	se_mean	sd	2.5%	97.5%	n_eff	Rhat
b1[1]	1.30	0.00580	0.250	0.87	1.8	1800	1
b1[2]	1.00	0.00068	0.030	0.95	1.1	1900	1

	mean	se_mean	sd	2.5%	97.5%	n_eff	Rhat
b1[3]	1.10	0.00120	0.061	0.98	1.2	2400	1
b1[4]	1.90	0.02100	0.930	0.55	4.1	1900	1
mu_c	0.71	0.00820	0.340	0.22	1.5	1700	1

Based on the above table we have obtained the posterior distribution for β_1 and there are four of these for each caste. We can see the 95% posterior interval for the parameters. We see the model converges well as the Rhat values are very close to 1.

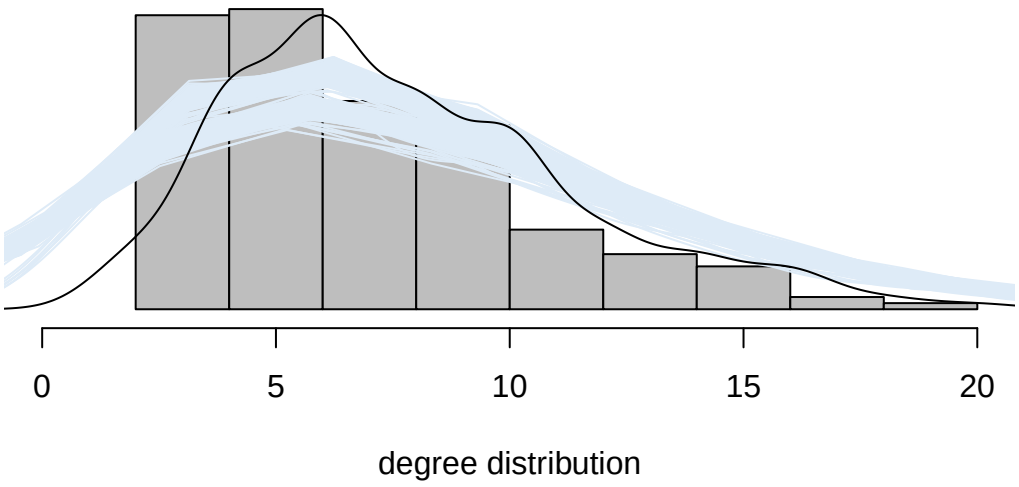
I plot the marginal posterior distributions of the parameters below.



- We estimate 4 values once again, for each caste under considertaion. This time we see that the ST coefficient is the highest followed by the general coefficient.
- This is a different result than with the normal model where the general coefficient was clearly higher. However, my suspicion is that this purely due to the fact that we have just one data point for ST nodes and the large error bars also suggest that this estimate is not stable. It is still surprising that the general coefficient is higher than the other two ("OBC" and "SC"), somewhat corroborating the result from the normal model.
- This is merely an introductory exercise and a pattern worth examining in closer detail.

Posterior predictive checking

I now plot the posterior predctive checks for the Poisson model with caste effects. Using the Poisson model I predict the degree distribution for the borrowing network and compare it to the raw data. I plot the histogram of the raw data. The light blue lines are simulated values for the dependent variable (degree of nodes in the borrowing network). The black line is the true density distribution of the dependent variable.



- We see that the Poisson distribution captures the positively skewed nature of the distribution. The Poisson model slightly underestimates the density at the mode but has aptly caught and predicted the long right tail. The Poisson model performs better than the normal model in that it captures the central tendency of the degree distribution in borrowing networks as well as the spread of values.

Conclusions

- We observe that degree centrality in the social engagement network is a good predictor of degree centrality in the borrowing network. Social engagement and access to credit are positively related. This is an interesting result and allows the researcher to delve deeper into the dynamics of social engagement in order to introduce social interventions that will aid access to credit in villages.
- We see that the Poisson model fits the data better than the normal model and predicts the data well. The Poisson model better captures the positively skewed nature of the data and the long right hand tail of the degree distribution in borrowing networks. It slighlty underpredicts the density at the mode.
- Importantly, we find out that belonging to the "general" caste usually means more access to credit, for a given level of social engagement. In the normal model we find that the "general" coefficient is the highest and in the Poisson model we find that it is second highest, although the "ST" coefficient has very high uncertainty in both cases because there is only one "ST" node in the village.
- As standard practice in Bayesian studies, we must leave the model open for expansion and checking. We performed some posterior predictive checks here but we could potentially expand these models, try different parametrizations and predictors and update our results.
- Criticisms we can offer for the models are:
  - network statistics other than "degree" have not been examined and this ignores a lot of information. Importantly "assortativity", "homophily" and "between-ness centrality" should be explored more carefully.

- We have to include many other covariates like religion, and other demographic data and understand if these reults still persist.
- The missing data also should be accounted for as a lot of information is thrown out when we drop nodes.
- This is a study of only one village and the focus is on developing a method that can be scaled to other villages. Scaling this method to other village networks is the next step to ensure generalizability of reults. The multilevel nature of the model should be expanded to incorporate the other villages in the network dataset into the study in a meaningful way.

References

Arora, Saurabh, and Bulat Sanditov. 2015. "Cultures of Caste and Rural Development in the Social Network of a South Indian Village." *Sage Open* 5 (3): 2158244015598813.

Betancourt, Michael. 2017. "A Conceptual Introduction to Hamiltonian Monte Carlo." *arXiv Preprint arXiv:1701.02434*.

Carpenter, Bob, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1).

Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. 2013. *Bayesian Data Analysis*. Chapman; Hall/CRC.

Gelman, Andrew, and others. 2006. "Prior Distributions for Variance Parameters in Hierarchical Models (Comment on Article by Browne and Draper)." *Bayesian Analysis* 1 (3): 515–34.

Gelman, Andrew, Daniel Simpson, and Michael Betancourt. 2017. "The Prior Can Often Only Be Understood in the Context of the Likelihood." *Entropy* 19 (10): 555.

Jackson, Matthew O. 2014. "Networks in the Understanding of Economic Behaviors." *Journal of Economic Perspectives* 28 (4): 3–22.