Caste, religion and access to credit in Karnataka's villages

Advait Rajagopal

December 18, 2017

Motivation

- The problem

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 set up a hierarchical Bayesian model that studies access to credit as a function of social engagement, while controlling for node level caste and religion effects. The dependent variable is the degree 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?"]

- The dataset

The network dataset is compiled by "Abdul Latif Jameel Poverty Action Lab" that is located in the economics department at MIT. 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 information about 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.

I have arbitrarily selected village 74 and perform some exploratory data analysis, visualize some data and patterns to highlight the importance of taking into account caste and religion effects on borrowing and finally set up two Bayesian hierarchical models (one Normal and one Poisson) to explain the data generating process. I also explain the reasons for my choices later on in this document.

Exploratory data analysis

I load the necessary packages and libraries first.

```
#rm(list = ls())
library(statnet)
library(foreign)
library(rstan)
rstan_options(auto_write = TRUE)
options(mc.cores = parallel::detectCores())
```

```
library(broom)
####
```

Now I import the necessary datasets.

```
#############
#village 74
#############
#import the ids of individuals in village 74
ind_ids_74 <- read.csv("/Users/Advait/Desktop/New_School/Fall17/Network_data/village_data/datav4.0/Data/

#Borrowing Money
bor_money_74 <- read.csv("/Users/Advait/Desktop/New_School/Fall17/Network_data/village_data/datav4.0/Da

#Social Engagement
social_eng_74 <- read.csv("/Users/Advait/Desktop/New_School/Fall17/Network_data/village_data/datav4.0/Da


#Remember torop rows and columns for nodes that dont have covariate information!
#import covariate info
#give the vertices attributes, tricky because of missing data
indiv_data <- read.dta("/Users/Advait/Desktop/New_School/Fall17/Network_data/village_data/datav4.0/Data/
```

I want to see how many people there are with covariate information in village 74.

```
#look at their ids
good_people_74 <- indiv_data$pid[indiv_data$village == 74]
length(good_people_74 )
```

```
## [1] 193
```

I drop the nodes with missing information (by dropping their corresponding rows and columns in the adjacency matrix to preserve its "squareness").

```
###I am starting over because of this data. tie individual ids to the adjacency matrix and drop those r
ind_ids_74$integer_ident <- seq(1:743)
#relevant people are ind_ids_74 with the missing data dropped
# ind_ids_74 <- ind_ids_74[,1:2]
good_ind_id_74 <- ind_ids_74[ind_ids_74$V1 %in% good_people_74,]

#relevant BORROW nodes
test1 <- as.matrix(bor_money_74)
test1 <- test1[good_ind_id_74$integer_ident,good_ind_id_74$integer_ident]
dim(test1)
```

```
## [1] 193 193
```

```
#relevant SOCIAL_ENG nodes
test2 <- as.matrix(social_eng_74)
test2 <- test2[good_ind_id_74$integer_ident,good_ind_id_74$integer_ident]
dim(test2)
```

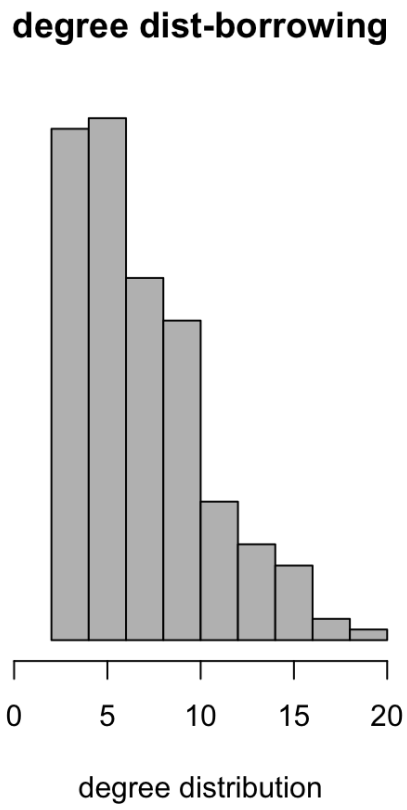
```
## [1] 193 193
```

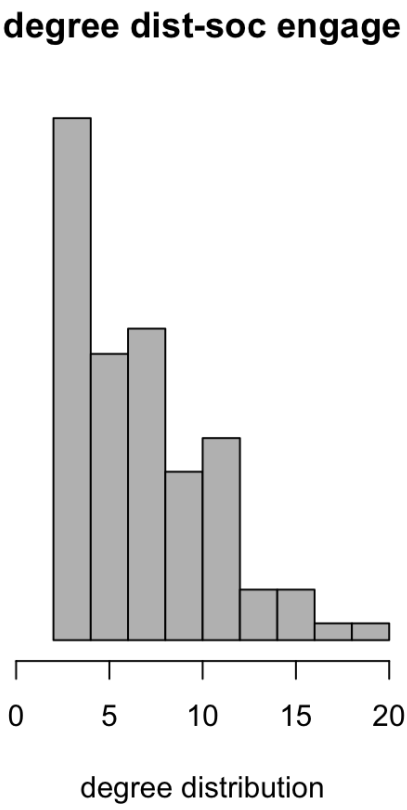
I then make the necessary network objects.

```
#Make network objects
#borowing
net_good_bor_74 <- as.network(x = test1,
```

```
                                directed = FALSE,
                                loops = FALSE,
                                matrix.type = "adjacency")
#social eng
net_good_soc_74 <- as.network(x = test2,
                                directed = FALSE,
                                loops = FALSE,
                                matrix.type = "adjacency")
```

Below we can see the degree distributions of these networks. The degree distribution appears to be right skewed with very few popular nodes and most nodes with a low degree.





Now I examine the covariate information and add the "caste" and "religion" information as node level covariates to the network object.

```
#EXPLORE attributes
#1. caste
caste_74 <- indiv_data$caste[indiv_data$village == 74]
length(caste_74)
```

```
## [1] 193
```

```
set.vertex.attribute(net_good_bor_74,
                     "caste",
                     as.character(caste_74))
set.vertex.attribute(net_good_soc_74,
                     "caste",
```

```
                    as.character(caste_74))
#2. religion
religion_74 <- indiv_data$religion[indiv_data$village == 74]
set.vertex.attribute(net_good_bor_74,
                    "religion",
                    as.character(religion_74))
set.vertex.attribute(net_good_soc_74,
                    "religion",
                    as.character(religion_74))
```

```
#summarize networks
summary.network(net_good_bor_74,
                print.adj = FALSE)
```

```
## Network attributes:
##   vertices = 193
##   directed = FALSE
##   hyper = FALSE
##   loops = FALSE
##   multiple = FALSE
##   bipartite = FALSE
##  total edges = 372
##    missing edges = 0
##    non-missing edges = 372
##  density = 0.02007772
##
## Vertex attributes:
##
##  caste:
##    character valued attribute
##    attribute summary:
##         GENERAL             OBC SCHEDULED CASTE SCHEDULED TRIBE
##               3             147              42               1
##
##  religion:
##    character valued attribute
##    attribute summary:
## HINDUISM    ISLAM
##      161       32
##   vertex.names:
##    character valued attribute
##    193 valid vertex names
##
## No edge attributes
```

```
summary.network(net_good_soc_74,
                print.adj = FALSE)
```

```
## Network attributes:
##   vertices = 193
##   directed = FALSE
##   hyper = FALSE
##   loops = FALSE
##   multiple = FALSE
##   bipartite = FALSE
```

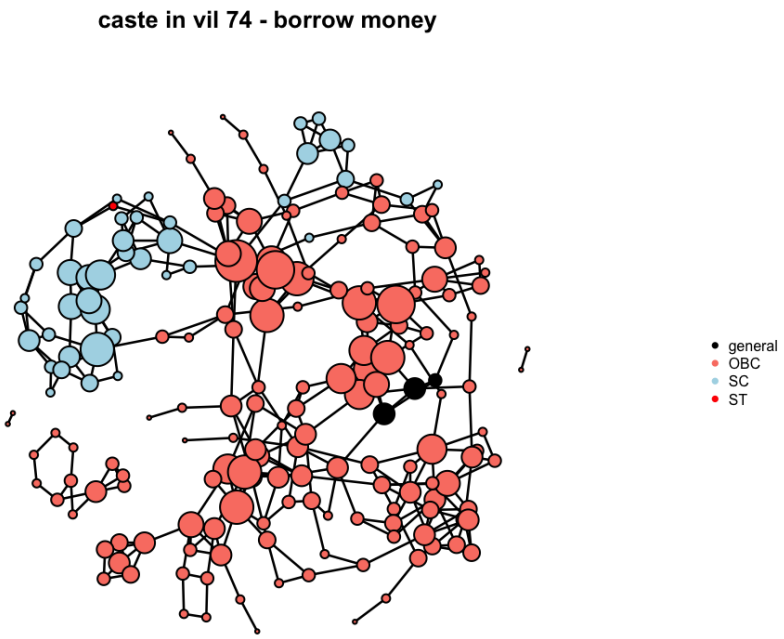
```
##  total edges = 362
##    missing edges = 0
##    non-missing edges = 362
##  density = 0.019538
##
## Vertex attributes:
##
##  caste:
##    character valued attribute
##    attribute summary:
##         GENERAL             OBC SCHEDULED CASTE SCHEDULED TRIBE
##               3             147              42               1
##
##  religion:
##    character valued attribute
##    attribute summary:
## HINDUISM    ISLAM
##      161       32
##   vertex.names:
##    character valued attribute
##    193 valid vertex names
##
## No edge attributes
```

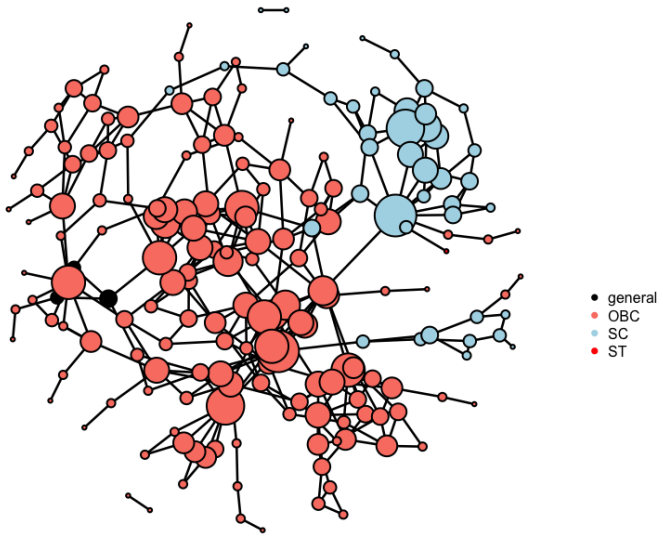
I prepare the networks for plotting by adding atrribute based colors.

```
#add colors based on attributes
#####
#Caste
#####
num_nodes <- 193
node_colors <- rep("",num_nodes)
for(i in 1:num_nodes){
  if(get.node.attr(net_good_bor_74,"caste")[i] == "GENERAL"){
    node_colors[i] <- "black"
  }
  if(get.node.attr(net_good_bor_74,"caste")[i] == "OBC"){
    node_colors[i] <- "salmon"
  }
  if(get.node.attr(net_good_bor_74,"caste")[i] == "SCHEDULED CASTE"){
    node_colors[i] <- "lightblue"
  }
  if(get.node.attr(net_good_bor_74,"caste")[i] == "SCHEDULED TRIBE"){
    node_colors[i] <- "red"
  }
}
num_nodes <- 193
node_colors <- rep("",num_nodes)
for(i in 1:num_nodes){
  if(get.node.attr(net_good_soc_74,"caste")[i] == "GENERAL"){
    node_colors[i] <- "black"
  }
  if(get.node.attr(net_good_soc_74,"caste")[i] == "OBC"){
    node_colors[i] <- "salmon"
  }
```

```
  if(get.node.attr(net_good_soc_74,"caste")[i] == "SCHEDULED CASTE"){
    node_colors[i] <- "lightblue"
  }
  if(get.node.attr(net_good_soc_74,"caste")[i] == "SCHEDULED TRIBE"){
    node_colors[i] <- "red"
  }
}
```

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 and SC nodes with very few general and just one ST 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.

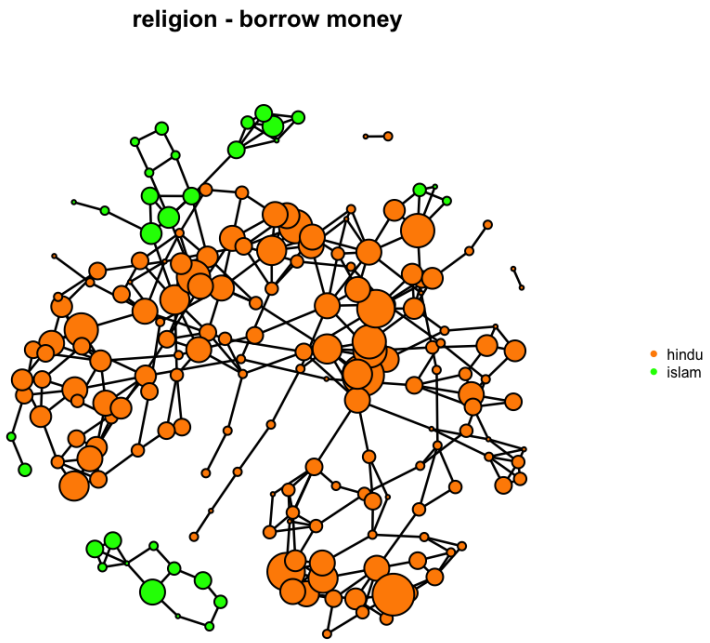


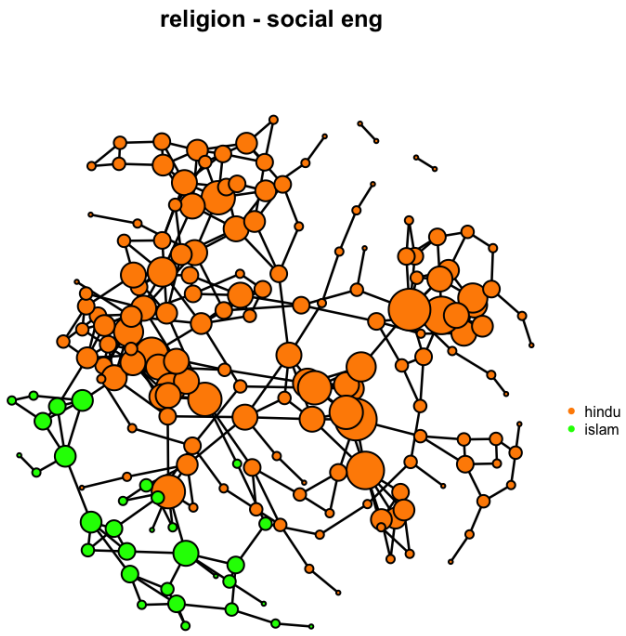


The “general” nodes are black, “OBC” are salmon, “SC” are light blue and “ST” is red.

```
#####
#Religion
#####
num_nodes <- 193
node_colors <- rep("",num_nodes)
for(i in 1:num_nodes){
  if(get.node.attr(net_good_bor_74,"religion")[i] == "HINDUISM"){
    node_colors[i] <- "darkorange"
  }
  if(get.node.attr(net_good_bor_74,"religion")[i] == "ISLAM"){
    node_colors[i] <- "green"
  }
}
num_nodes <- 193
node_colors <- rep("",num_nodes)
for(i in 1:num_nodes){
  if(get.node.attr(net_good_soc_74,"religion")[i] == "HINDUISM"){
    node_colors[i] <- "darkorange"
  }
  if(get.node.attr(net_good_soc_74,"religion")[i] == "ISLAM"){
    node_colors[i] <- "green"
  }
}
```

I then plot the same networks this time showing the religion of the nodes. There are more “hindu” nodes than “islam nodes”. We see that the muslim community appears to be at the periphery of the network and a large section borrrows money only internally. Their degrees are relatively lower in the borrowing network and the social engagement network.





The "hindu" nodes are orange and the "islam" nodes are in green. Adding legends to network plots with user defined `par` specifications has been tricky.

Hierarchical Models

I set up a normal model and a Poisson model to capture the relationship between social engagement and access to credit.

The normal model fits a hierarchical linear regression and is useful because it has easily interpretable coefficients. Each coefficient can be seen as the marginal impact on the outcome variable.

The Poisson model is appropriate because it is a discrete distribution for count data and given that we are modeling degree distributions this seems more correct. The support of the normal distribution extends into the negative half of the real line and this is not right for this model. The Poisson model performs better than the Normal model, but the Normal model is easy to interpret, so I leave them both in.

I first consider "caste" level effects and then "religion" level effects. It might be possible to build both effects into one model but I have not done this for now as I would prefer to interpret these coefficients separately. I use Stan 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) to accomplish this. The posterior distribution is proportional to the product of the prior and likelihood.

Normal model for caste level effects

The full model is described in the equations below;

I fit a normal model where β_1 is the slope coefficient that measures the caste specific marginal impact of social engagement degree on the degree in the borrowing network. I use normal priors on the parameters β_0, β_1 and μ_c. I use 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.

```
#model preparation
deg_74 <- degree(net_good_bor_74)
deg_social_eng <- degree(net_good_soc_74)
rel_74 <- as.numeric(as.factor(get.node.attr(net_good_bor_74,"religion")))
cas_74 <- as.numeric(as.factor(get.node.attr(net_good_bor_74,"caste")))
N <- length(deg_74)
#marginal posteriors
c_vec <- c("gen","obc","sc","st")
r_vec <- c("hinduism","islam")
```

```
#normal model for caste
apsta_4 <- "
data{
  int <lower = 0> N;
  real deg_74[N];
  int cas_74[N];
  real deg_social_eng[N];
}
parameters{
  real b0;
  real b1[4];
  real <lower = 0> sigma;
  real mu_c;
  real <lower = 0> tau_c;
}
model{
  for(i in 1:N){
    deg_74[i] ~ normal(b0 + b1[cas_74[i]]*deg_social_eng[i], sigma);
  }
  b0 ~ normal(0,1);
  b1 ~ normal(mu_c,tau_c);
  mu_c ~ normal(0,0.5);
  tau_c ~ cauchy(0,10);
  sigma ~ cauchy(0,10);
}
generated quantities{
  real y_pred[N];
  for(i in 1:N){
    y_pred[i] = normal_rng(b0 + b1[cas_74[i]]*deg_social_eng[i], sigma);
  }
}
"
fit1 <- stan(model_code = apsta_4,
             data = list("N",
                         "deg_74",
```

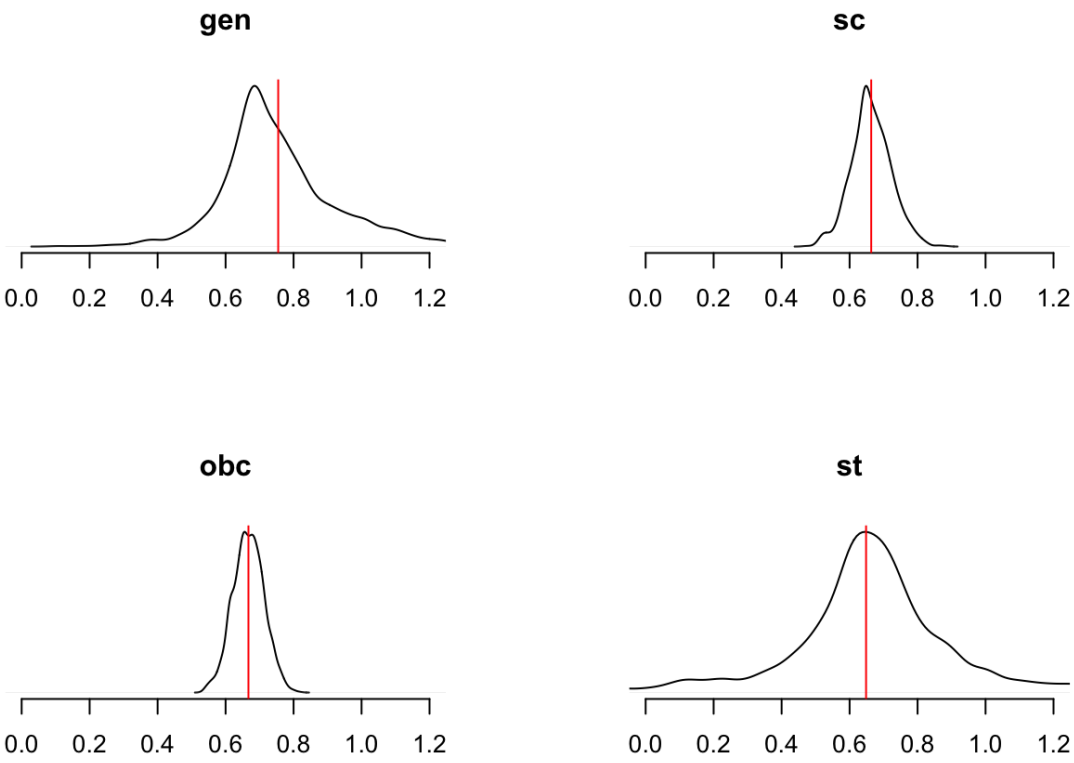
```
                          "cas_74",
                          "deg_social_eng"),
              iter = 1000,
              chains = 3)
```

```
ext1 <- extract(fit1)
print(fit1, pars = c("b0","b1","mu_c","tau_c"))
```

```
## Inference for Stan model: f787512a8b84f2aeaeb75c72cd5acb9d.
## 3 chains, each with iter=1000; warmup=500; thin=1;
## post-warmup draws per chain=500, total post-warmup draws=1500.
##
##       mean se_mean   sd  2.5%  25%  50%  75% 97.5% n_eff Rhat
## b0    2.66    0.03 0.39  1.95 2.36 2.67 2.93  3.41   224 1.03
## b1[1] 0.74    0.01 0.19  0.45 0.63 0.70 0.82  1.19   202 1.02
## b1[2] 0.66    0.00 0.05  0.57 0.63 0.66 0.69  0.76   273 1.01
## b1[3] 0.66    0.00 0.06  0.54 0.61 0.65 0.70  0.78   231 1.02
## b1[4] 0.61    0.01 0.34 -0.30 0.55 0.64 0.75  1.28   527 1.01
## mu_c  0.61    0.01 0.19  0.08 0.57 0.64 0.71  0.90   453 1.01
## tau_c 0.26    0.03 0.35  0.02 0.05 0.13 0.32  1.25   128 1.02
##
## Samples were drawn using NUTS(diag_e) at Thu Sep 12 22:47:46 2019.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

We can see the posterior intervals of the β_1 parameters and there are 4 (one for each caste). $\beta_1, \beta_2, \beta_3, \beta_4$ correspond 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 could mean that historically "upper" caste, despite being a minority is wealthier and has a lot of 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. 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.



I now do some posterior predictive checks. Using my model I predict the degree distribution for the borrowing network and compare it to the raw data. For perfect predictions all the points should cluster around the 45 degree line.

Poisson model for caste level effects

The full model is described in the equations below;

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 data. I change the likelihood functions but the priors on the parameters remain the same.

```
#normal model for caste
apsta_5 <- "
data{
  int <lower = 0> N;
  int deg_74[N];
  int cas_74[N];
  int deg_social_eng[N];
}
parameters{
  real b1[4];
  real mu_c;
  real <lower = 0> tau_c;
}
```

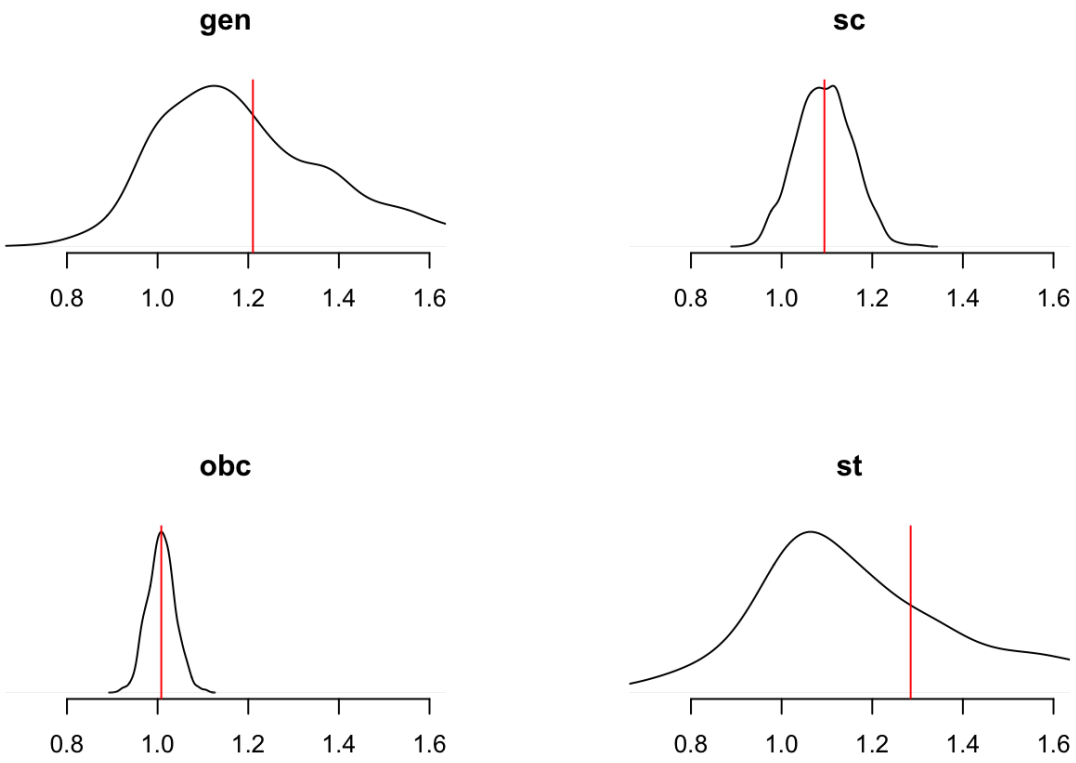
```
model{
  for(i in 1:N){
    deg_74[i] ~ poisson(b1[cas_74[i]]*deg_social_eng[i]);
  }
  b1 ~ normal(mu_c,tau_c);
  mu_c ~ normal(1,1);
  tau_c ~ cauchy(0,10);
}
generated quantities{
  real y_pred[N];
  for(i in 1:N){
    y_pred[i] = poisson_rng(b1[cas_74[i]]*deg_social_eng[i]);
  }
}
"
fit2 <- stan(model_code = apsta_5,
             data = list("N",
                         "deg_74",
                         "cas_74",
                         "deg_social_eng"),
             iter = 1000,
             chains = 3)
```

```
ext2 <- extract(fit2)
print(fit2, pars = c("b1","mu_c","tau_c"))
```

```
## Inference for Stan model: 9e65f2bef1deb013b0bd2cf2bd82e45c.
## 3 chains, each with iter=1000; warmup=500; thin=1;
## post-warmup draws per chain=500, total post-warmup draws=1500.
##
##       mean se_mean   sd 2.5%  25%  50%  75% 97.5% n_eff Rhat
## b1[1] 1.21    0.01 0.20 0.90 1.08 1.17 1.31  1.69   549 1.00
## b1[2] 1.01    0.00 0.03 0.95 0.99 1.01 1.03  1.07   179 1.00
## b1[3] 1.10    0.00 0.06 0.99 1.06 1.10 1.13  1.21   500 1.01
## b1[4] 1.27    0.02 0.40 0.74 1.05 1.16 1.38  2.38   447 1.00
## mu_c  1.13    0.01 0.27 0.63 1.03 1.10 1.21  1.80   443 1.01
## tau_c 0.38    0.04 0.60 0.05 0.10 0.20 0.40  1.94   247 1.00
##
## Samples were drawn using NUTS(diag_e) at Thu Sep 12 22:48:32 2019.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

We see that the model coverges well. Again the highest β_1 value is for the "general" caste, meaning that this has the greatest marginal impact on the expected value of the Poisson distribution.

I plot the marginal posterior distributions of the parameters below.



I now plot the posterior predctive checks for the poisson model with caste effects.

Normal model for religion level effects

The full model is given below; β_1 is the slope coefficient that measures the religion specific marginal impact of social engagement degree on the degree in the borrowing network. There are two religions, "hinduism" and "islam".

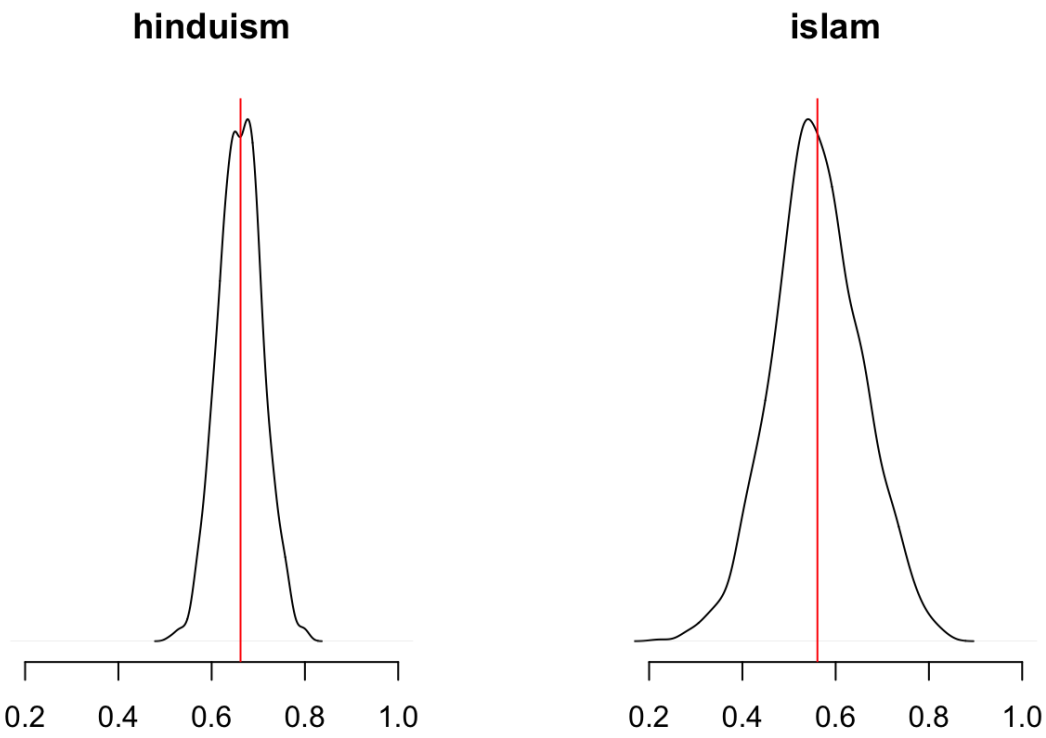
```
#normal model for religion
apsta_6 <- "
data{
  int <lower = 0> N;
  real deg_74[N];
  int rel_74[N];
  real deg_social_eng[N];
}
parameters{
  real b0;
  real b1[2];
  real <lower = 0> sigma;
  real mu_r;
  real <lower = 0> tau_r;
}
model{
```

```
  for(i in 1:N){
    deg_74[i] ~ normal(b0 + b1[rel_74[i]]*deg_social_eng[i], sigma);
  }
  b0 ~ normal(0,1);
  b1 ~ normal(mu_r,tau_r);
  mu_r ~ normal(0,0.5);
  tau_r ~ cauchy(0,10);
  sigma ~ cauchy(0,10);
}
generated quantities{
  real y_pred[N];
  for(i in 1:N){
    y_pred[i] = normal_rng(b0 + b1[rel_74[i]]*deg_social_eng[i], sigma);
  }
}
"
fit3 <- stan(model_code = apsta_6,
             data = list("N",
                         "deg_74",
                         "rel_74",
                         "deg_social_eng"),
             iter = 1000,
             chains = 3)
```

```
ext3 <- extract(fit3)
print(fit3, pars = c("b0","b1","mu_r","tau_r"))
```

```
## Inference for Stan model: b23c1806476102a3b3c8c46ff3b25d2a.
## 3 chains, each with iter=1000; warmup=500; thin=1;
## post-warmup draws per chain=500, total post-warmup draws=1500.
##
##       mean se_mean   sd  2.5%  25%  50%  75% 97.5% n_eff Rhat
## b0    2.64    0.04 0.43  1.93 2.32 2.65 2.96  3.48    92 1.02
## b1[1] 0.67    0.00 0.05  0.58 0.64 0.67 0.70  0.76   392 1.00
## b1[2] 0.58    0.01 0.09  0.39 0.52 0.58 0.65  0.75   199 1.02
## mu_r  0.42    0.02 0.37 -0.53 0.26 0.54 0.67  0.92   358 1.02
## tau_r 0.93    0.09 1.62  0.02 0.12 0.35 1.02  5.21   308 1.02
##
## Samples were drawn using NUTS(diag_e) at Thu Sep 12 22:49:19 2019.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

We see that the model coverges well. The coefficient on hindusim ($\beta_1[1]$) is higher indicating that being hindu predisposes a node to a higher degree in the borrowing network if the node is more socially engaged. I plot the marginal posterior distributions of the religion level slope parameters.



I plot the posterior predictive checks for the model.

Poisson model for religion level effects

The full model is described in the equations below;

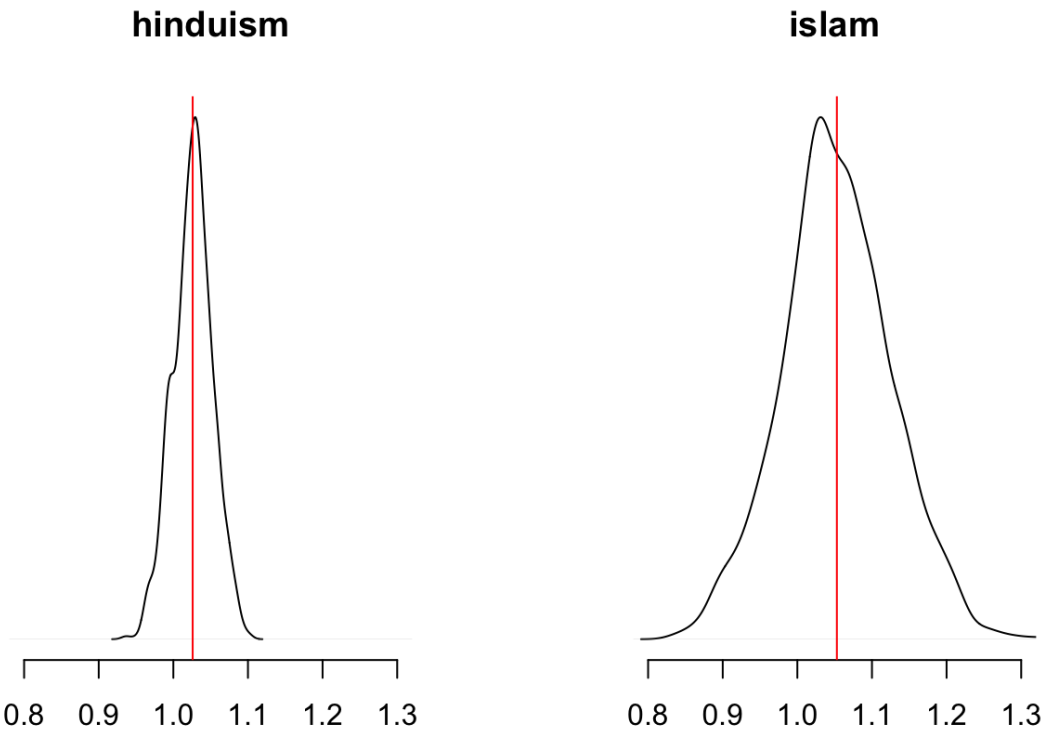
```
#poisson model for religion
apsta_7 <- "
data{
  int <lower = 0> N;
  int deg_74[N];
  int rel_74[N];
  int deg_social_eng[N];
}
parameters{
  real b1[2];
  real mu_r;
  real <lower = 0> tau_r;
}
model{
  for(i in 1:N){
    deg_74[i] ~ poisson(b1[rel_74[i]]*deg_social_eng[i]);
  }
  b1 ~ normal(mu_r,tau_r);
```

```
  mu_r ~ normal(1,0.5);
  tau_r ~ cauchy(0,10);
}
generated quantities{
  real y_pred[N];
  for(i in 1:N){
    y_pred[i] = poisson_rng(b1[rel_74[i]]*deg_social_eng[i]);
  }
}
"
fit4 <- stan(model_code = apsta_7,
             data = list("N",
                         "deg_74",
                         "rel_74",
                         "deg_social_eng"),
             iter = 1000,
             chains = 3)
```

```
ext4 <- extract(fit4)
print(fit4, pars = c("b1","mu_r","tau_r"))
```

```
## Inference for Stan model: 571f92a668d39b7026b93659e7bb5a02.
## 3 chains, each with iter=1000; warmup=500; thin=1;
## post-warmup draws per chain=500, total post-warmup draws=1500.
##
##       mean se_mean   sd 2.5%  25%  50%  75% 97.5% n_eff Rhat
## b1[1] 1.03    0.00 0.03 0.97 1.01 1.03 1.04  1.09   491 1.01
## b1[2] 1.05    0.00 0.07 0.91 1.00 1.05 1.10  1.20   809 1.00
## mu_r  1.05    0.01 0.25 0.53 0.96 1.04 1.13  1.68   437 1.00
## tau_r 0.68    0.06 1.37 0.03 0.09 0.23 0.63  4.34   468 1.00
##
## Samples were drawn using NUTS(diag_e) at Thu Sep 12 22:50:05 2019.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

I plot the marginal posterior distributions of the religion level parameters.



Conclusions

- We observe that degree centrality in the social engagement network is a good predictor of degree centrality in the borrowing network. Thus being more engaged predisposes an individual to have more access to credit. This is an interesting result and allows the researcher to delve deeper into the dynamics of social engagement in order to introduce social interventions etc.
- We see that the Poisson model really fits the data well and predicts well.
- We find out that belonging to the "general" caste usually means more access to credit.
- There is ambiguity in the religion models as "hinduism" has greater marginal effects in the normal model but "islam" does in the Poisson model.
- 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 that other network statistics have not been examined and this ignores a lot of information. Importantly "assortativity" should be explored more carefully. The missing data also should be accounted for as a lot of information is thrown out when we drop nodes. This is also a study of only one village and almost no generalizations can be made to how communities behave in other villages and how caste and religion dynamics impact the village. Importantly, the multilevel nature of the model should be expanded to incorporate the other villages in the network data into the study in a meaningful way.