Modeling Strategic Information Sharing in Indian Villages

Jeff Jacobs
Stanford University
jjacobs3@stanford.edu

Arun Chandrasekhar
Stanford University
arungc@stanford.edu

Matthew O. Jackson
Stanford University
jacksonm@stanford.edu

Emily Breza
Columbia University
ebreza@columbia.edu

December 13, 2013

Abstract

Several recent research initiatives in the field of Development Economics have studied and provided substantial insight into the mechanics of information diffusal in small village social networks, in particular those without formal mass communication infrastructure in place (*i.e.*, email, television). This research aims to further explore the dynamics of communication in these networks, focusing on the possibility that village households share and withold information *strategically*, only sharing with neighbors exhibiting certain characteristics, in an individually rational manner. A model of probabilistic information diffusion is developed, the parameters of which are estimated via the Generalized Method of Moments (GMM) and tested with empirical data.

1 Background: Information Diffusion in Village Networks

Banerjee, et al. (2013) studied the way in which the spread of information about microfinance programs is affected by whom in the community the information is given to initially (the "information injection point"), finding that giving the information to more central figures in the community significantly increases the diffusal of the information throughout the village [1]. Conley and Udry (2008) studied the adoption of new agricultural technology in Ghana from a social network standpoint, and concluded that "information has value in these villages, as do the network connections through which that information flows" [4]. Most recently, Breza and Chandrasekhar (2013) explored the possibility that microloan recipients in rural villages of Karnataka, India can be influenced to save and invest these loans more efficiently if they are paired with a peer monitor who is influential in the village, and indeed found that "randomly assigned central monitors generate considerably larger increases in savings attainment" [2].

2 Generalized Method of Moments Estimation

Before presenting the model, we describe the Generalized Method of Moments (GMM) Estimation framework, as this is the crucial statistical (and CS229-related) tool used in this work. GMM is a generalization of Maximum Likelihood Estimation, which is used primarily in econometrics to learn optimal model parameter estimates when the probability distribution of the data is not fully specified or completely unknown and thus MLE is not feasible. We will see, however, that if this distribution is specified we can recover the maximum likelihood estimator from within the GMM framework.

The efficacy of a particular GMM estimation depends upon the specification of informative *moment conditions*, characteristics of the data which will be used to "match up" the data generated by the model with empirical data. Specifically, the Euclidean distance between the sample (empirical) moments and the theoretical (model-generated) moments will be minimized to obtain the optimal model parameters.

As an example, consider the case of standard linear regression, in which we try to estimate the unknown parameters θ which generate the observed parameters y as a function of observed data x with stochastic noise u: $y = \theta^T x + u$. The standard linear regression model asserts that, conditioned on the observations x, the noise terms have mean zero, *i.e.*that $E[u|x] = 0$. This condition, along with the regression equation, leads to the fact that $E[xu] = E[x(y - \theta^T x)] = 0$, which we use as our *moment condition.* Thus we would like to find an estimate $\widehat{\theta}$ of the true value of θ, such that (1) holds for some N samples $\{(x^{(i)}, y^{(i)})\}_{i=1}^N$ we've obtained (we don't have access to the true data, otherwise we could just solve for θ directly). To achieve this, the GMM "trick" is to rewrite (1) in terms of our sample data, by replacing all of the expected values with sample means:

$$\frac{1}{N}\sum_{i=1}^{N} x^{(i)}\hat{u}^{(i)} = \frac{1}{N}\sum_{i=1}^{N} x^{(i)}(y^{(i)} - \widehat{\theta}^T x^{(i)}) = 0.$$

Now we simply solve for the optimal parameter estimate $\widehat{\theta}$ and obtain:

$$\widehat{\theta} = \left(\sum_{i=1}^{N} x^{(i)}(x^{(i)})^T\right)^{-1} \sum_{i=1}^{N} x^{(i)}y^{(i)} = (X^T X)^{-1}X^T y,$$

which we recognize as our standard OLS estimator.

The power of GMM is further revealed by noticing that the Maximum Likelihood Estimator can be obtained by noting that the true θ which maximizes the data's likelihood satisfies the condition $E\left[\nabla_\theta \ell(\theta \mid x)\right] = 0$, thus via the GMM "trick" we again rewrite this condition in terms of a given sample of size N as

$$\frac{1}{N}\sum_{i=1}^{N}\left[\nabla_\theta \ell(\widehat{\theta} \mid x_i)\right] = 0,$$

which, when solved, results in precisely the Maximum Likelihood Estimator $\hat{\theta}$ of the data.

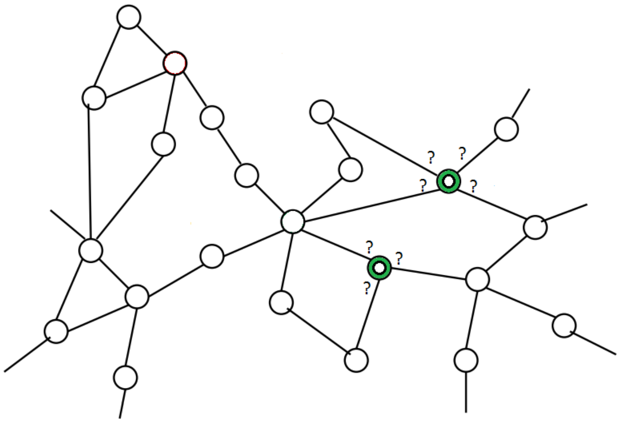
3 The Model

We now return to the model of strategic information. We have network data from 75 villages in rural Karnataka, India. In 39 villages, we had randomly invited individuals to come to participate in a game. We had told about 10 people about this at random and then asked them to invite their friends, telling them that there were only 20 slots but that these are cooperative games with others and therefore they may face better payoffs if they bring friends. Thus the core idea of the model is that there is a tension in the passing of information: if you don't tell others, then you don't get to play with your friends (which generates higher payoffs), but if you do tell others, they may tell others, which generates low returns for you (since there is more rivalry). Therefore, our model implements cutoff rules: namely, that you only tell neighbors that themselves are not too popular, so they do not pass information too much. Formally, we define the diffusion model in terms of 3 parameters:

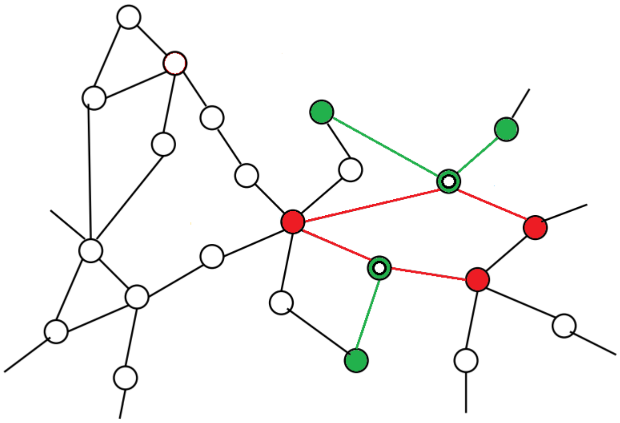
- $\epsilon \in [0, 1]$: the probability that you accidentally inform someone you did not intend to,
- $q \in [0, 1]$: the probability that you inform someone that you intended to,
- $\tau \in [0, 1]$: the cutoff parameter, and
- $p \in [0, 1]$: the probability that you act on a given piece of information,

where the cutoff parameter asserts that you inform all neighbors $j \in N_i$ if $p(d_j) \leq \tau$, where $p(\cdot)$ is the percentile function from the network's global degree distribution.

The model's dynamics are illustrated in the following figures – first, invites are given to randomly selected household representatives (the nodes containing white dots), who then are faced with the strategic decision of whether to invite their neighbors:



Households then choose to share with neighbors whose degree is below a certain threshold determined by *tau* (green edges indicate that information was shared, red edges indicate that information was not shared):



In addition to the diffusion model, however, we need to model the relationship between how informed a given household is and the probability that someone in the household actually shows up for the game, so that the empirical data can be compared with the simulated data via GMM. To incorporate this into the overall model, we introduce the following notation:

Let $\mathbf{Q}(\theta)$ be a matrix of probabilities, defined as

$$\mathbf{Q}_{ij} = \begin{cases} q + \epsilon & \text{if } p(d_j) \leq \tau \\ \epsilon & \text{if } p(d_j) > \tau \end{cases}.$$

Then we can define another matrix $\mathbf{M}_T$ such that

$$\mathbf{M}_T(\theta) := \sum_{t=1}^{T} (\mathbf{Q}(\theta) \circ \mathbf{A}(G))^t,$$

where $\mathbf{A}(G)$ is just the incidence matrix of the graph representing the social network. The result of this definition is that $\mathbf{M}_T$ represents the expected number of times that information given to household i in the network reaches household j after T timesteps (in our simulations, we set $T = 3$). Then if we let m_i be the ith row sum of $\mathbf{M}_T$ we can model the probability $P(y_i = 1)$ of someone from a household i showing up as

$$P(y_i = 1|m_i) = 1 - (1 - p)^{m_i}.$$

Thus, we have that we need to estimate the parameter vector $\theta = (\epsilon, q, \tau, p)$ to fully specify our model. In the next section, we outline the procedure used to estimate these parameters via GMM.

4 GMM Objective Optimization

As we are using GMM to estimate the parameters from the prior section, it is crucial that we develop a set of 4 moments (thus fully specifying the system) so we can perform the "matching up" of population and sample moments explained in the GMM section. The moments we developed are as follows:

1. The share of invited households with high centrality that showed up
2. The share of invited households with low centrality that showed up
3. The share of non-invited households with high centrality that showed up
4. The shared of non-invited households with low centrality that showed up

For each of these, centrality is defined to be the eigenvector centrality of the household with respect to the village's adjacency matrix. MATLAB code (available upon request) was written to compute the empirical values of these moments, as well as code to simulate the model and compute their simulated values. The moments were plotted pairwise and no pairs were found to be strongly collinear.

Once these moments are defined and computed, we let $g_{sim,r}(\theta)$ represent the vector of sample moments of the simulation data for household r and let $g_{emp,r}$ represent the vector of sample moments of the empirical data, and obtain the optimal parameter settings $\widehat{\theta}$ via

$$\widehat{\theta} := \arg\min_{\theta} \left(\frac{1}{R} \sum_{r} g_{sim,r}(\theta) - g_{emp,r} \right)^T \left(\frac{1}{R} \sum_{r} g_{sim,r}(\theta) - g_{emp,r} \right).$$

It should be stressed here that the ability to obtain the optimal parameters settings by solving this equation represents the primary advantage of GMM over Maximum Likelihood estimation for our model – we cannot observe a household's θ directly, and GMM allows us to estimate it without making any prior assumptions about its distribution, as ML would require us to do in this case.

5 Results

The primary computational challenge presented by the model is that the objective function given in the prior section is not convex in its parameters, as can be seen by plotting it as a function of any of θ's elements with the other three held constant. Thus we implemented a simulated annealing procedure in MATLAB to search for its minima and attempt to find globally optimal solutions, which resulted in the parameter settings

$$q = 0.11, \epsilon = 0.03, \tau = 0.48, p = 0.18,$$

and in a GMM error (the value of the minimand) of 0.0395. We then performed a heuristic bootstrap procedure, following the method of [1], with $B = 1000$ village-level Bayesian bootstrap estimates and found that the parameter of interest τ is indeed statistically significant, with a standard error of 0.04.

We also measured the temporal robustness of the model by re-running the simulation with $T = 2$, representing the "myopic" case wherein only the invited households have a chance to share information, and with $T = 4$, representing the case where all households in each experimental village (with 3 exceptions) would have had a chance to receive the information. In both cases, we find that the resulting τ is within one standard error of the $T = 3$ case, indicating robustness to measurement over different time spans.

6 Conclusions/Next Steps

Given the statistically significant τ estimate, we have **evidence that village households indeed engage in strategic and individually rational decision making** when sharing information with neighbors. This is of particular interest to policymakers, as several financial instruments depend heavily on participants' strategic divulgence or withholding of information. For example, consider the standard microfinance model in which loans are given out to groups of individuals, and if one of these individuals does not repay the loan it is up to the remaining group members to make up the difference or default on the loan. It is thus of crucial importance to an individual to choose group members whom they believe are trustworthy and capable of repaying, inducing a strategic decision-making process in which an individual must examine the characteristics of their neighbors and decide whether to invite them to their loan group or not. Our model represents a special case of this process, wherein the utilized "strategic filtering" characteristic is the percentile of the neighbor's degree.

This observation indicates a natural next step for our research, which is to further study the dynamics of information diffusal in the village networks using different characteristics as the basis for the strategic choice. We have already begun this study, by constructing demographic vectors $x_{r,i}$ representing demographic information about the household i in village r such as reported income, caste, religious affiliation, roofing material of their home, type of latrine, access to electricity, and so on. We are currently developing a regression procedure to estimate the effect changes in any of these variables have on neighbors' information sharing choices, which will allow greater insight into the dynamics of information diffusal in village networks.

References

[1] Banerjee, A., Chandrasekhar, A., Duflo, E., and Jackson, M. 2013. "The Diffusion of Microfinance." *Science*, 341.

[2] Breza, E. and Chandrasekhar, A. 2013 "Savings Monitors." *Working Paper.*

[3] Cole, S., Sampson, T., and Zia, B. 2011. "Prices or Knowledge? What Drives Demand for Financial Services in Emerging Markets?" *The Journal of Finance*, 66(6): 1844–1867.

[4] Conley, T. and Udry, C. 2008. "Learning About a New Technology: Pineapple in Ghana." *American Economic Review*, 100: 35–69.

[5] MacKinnon, J. 2006. "Bootstrap Methods in Econometrics." *The Economic Record*, 82: 2–18.

[6] Matyas, L. 2007. *Generalized Method of Moments Estimation.* Cambridge University Press.