MIS Quarterly

RESEARCH ARTICLE

IS A CORE–PERIPHERY NETWORK GOOD FOR KNOWLEDGE SHARING? A STRUCTURAL MODEL OF ENDOGENOUS NETWORK FORMATION ON A CROWDSOURCED CUSTOMER SUPPORT FORUM

Yingda Lu
Lally School of Management, Rensselaer Polytechnic Institute,
Troy, NY 12180 U.S.A. {luy6@rpi.edu}

Param Vir Singh
David A. Tepper School of Business, Carnegie Mellon University,
Pittsburgh, PA 15213 U.S.A. {psidhu@cmu.edu}

Baohong Sun
Cheung Kong Graduate School of Business, 1 E Chang'an Avenue, DongDan,
Dongcheng Qu, Beijing Shi CHINA {bhsun@ckgsb.edu.cn}

Appendix A

Estimation and Identification

Estimation and Initial Values

There are two main advantages of using oblivious equilibrium(OE) in our context. First of all, it has been shown that OE works well when there are a large number of agents in the model (Weintraub et al. 2010). Unlike the traditional context in economics where there are a handful of firms competing on the market, there are hundreds, if not thousands of individuals on our social media platform. The large number of agents in our context makes the exact computation of Markov Perfect Equilibrium (hereafter MPE) almost impossible. On the other hand, OE offers an alternative equilibrium strategy which is very close to MPE in such context with many agents, thus the performance of OE is particularly good for modeling individual decisions in our context. Second, OE also provides an appealing behavioral model that is also highly consistent with individual decision patterns in our setting. Because there are many users, it is unrealistic for individuals to keep track of the states of every user on the forum. Thus actual individual decision rules could be inconsistent with the assumptions in MPE where agents make decisions based on the states of every agent in the market. However, OE assumes that in a market with many agents, each individual makes nearly optimal decisions based on her own state as well as the long-run average market state, rather than keeping track of everyone's states. This is because the changes of individual states in each period can be averaged out due to the large number of agents (Weintraub et al. 2008). In other words, we approximate each individual's optimal decision by her behavior under the assumption that she only keeps track of long-term aggregate state level on the platform. We use $\mathbb{Z}$ to denote the expected average state of the market in the long run, which includes market average knowledge level, social status level, and the distribution of unobserved individual type. Hereafter, we use $\boldsymbol{\sigma}_i$ to denote OE strategy to ask and answer question as a function of her own state variables and the private shock for individual i: $\boldsymbol{\sigma}_i : \boldsymbol{h}_i \times \mathbb{Z} \times \boldsymbol{\varepsilon}_i, \leftrightarrow \boldsymbol{A}_i$ where A_i is the set of all actions individual i can take. An individual also expects her peers to use the publicly known OE strategy to make their own decisions. We use $\boldsymbol{\sigma}_i'$ to denote the OE strategy for individuals other than i. Because individual decisions depend only on their state levels in oblivious equilibrium, the oblivious value function can be defined as

$$\overline{V}_l\left(\boldsymbol{h}_{i,t}|\boldsymbol{\sigma}_i',\boldsymbol{\sigma}_i\right)=E\left[\left[\sum_{\tau=t}^{\infty}\beta^{\tau-t}U_i(\tau)|\boldsymbol{h}_{i,t}\right]\right]$$

Here, the oblivious value function $\overline{V}_l\left(\boldsymbol{h}_{i,t}|\boldsymbol{\sigma}_i',\boldsymbol{\sigma}_i\right)$ is the expected net present value of an individual with state $\boldsymbol{h}_{i,t}$ and follows oblivious strategy $\boldsymbol{\sigma}_i'$, when her peers with long-term state $\mathbb{Z}$ follow strategy $\boldsymbol{\sigma}_i$. Then a strategy profile $\boldsymbol{\sigma}^* = \{\boldsymbol{\sigma}_1^*, \ldots, \boldsymbol{\sigma}_N^*\}$ is an OE strategy for a market with a large amount of agents, in which this strategy optimizes an oblivious value function

$$sup_{\boldsymbol{\sigma}_i'}\overline{V}_l\left(\boldsymbol{h}_{i,t}|\boldsymbol{\sigma}_i^*,\boldsymbol{\sigma}_i\right)=\overline{V}_l\left(\boldsymbol{h}_{i,t}|\sigma_i,\boldsymbol{\sigma}_i\right),\forall i,\mathscr{H}$$

One shortcoming of OE is that we cannot explicitly model the dyadic interaction between individuals whether they answer a question from their peers. This is because OE assumes that agents only take into account aggregate state of peers and that agents play long-run equilibrium strategies, while the nature of dyadic interaction requires individuals to keep track of the states of all users of the platform. Thus we discretize state space, and adapt individual decision of answering question in equation (2)

$$s_{iht}=\begin{cases}k, individual\ i\ answers\ k\ questions\ from\ peers\ with\ state\ level\ h\\ \vdots\\ 2, if\ individual\ i\ answers\ 2\ questions\ from\ peers\ with\ state\ level\ h\\ k, if\ individual\ i\ answers\ 1\ question\ from\ peers\ with\ state\ level\ h\end{cases}\tag{2'}$$

As we can see from equation (2'), neither asking nor answering decisions will require individuals to track the state of their peers. Instead, individuals only keep track of the proportion of customers in each social status level. Because the social media platform is in steady state, the number of questions asked by customers in each state level should also be steady across periods. Thus policy functions derived from decision rule in equation (2') are expected to be very similar to that derived from equation (2).

The time line of the decisions at time t is as follows:

(1) Everyone observes their own states and the states of everyone else in the community.

(2) Everyone receives their private shocks for the decision of asking question.

(3) Everyone makes predictions of their peers' decisions based on equilibrium strategy given their information on others' states in current period, and using this prediction everyone simultaneously makes decisions on whether they are going to ask a question.

(4) Everyone observes the questions asked (i.e., they know who asked questions in current period).

(5) Everyone receives their private shocks for the decision of answering questions.

(6) Given the information on who asked questions in the current period and the predictions on others' decisions of answering questions, everyone simultaneously makes decisions on whether to provide an answer for each of the questions proposed.

(7) The state variables, accumulated knowledge K_{it} and social status R_{it} are updated.

We adapt oblivious equilibrium (Weintraub et al. 2010) to our dynamic knowledge sharing framework. We use $\boldsymbol{\theta}^0$ and $\boldsymbol{\pi}^0$ to denote initial guess of structural parameters and distribution of unobserved individual heterogeneity respectively. Given $\boldsymbol{\theta}^m$ and $\boldsymbol{\pi}^m$ in the m^{th} iteration, we can update the $(m + 1)^{th}$ iteration as

(1) Calculate the optimal asking and answering decisions $\boldsymbol{\sigma}^{m+1}$ that maximize the value function $V(\boldsymbol{s}|\boldsymbol{\sigma}^{m+1}, \boldsymbol{\sigma}^m)$

(2) $\delta = ||\boldsymbol{\sigma}^{m+1} - \boldsymbol{\sigma}^m||$, $m = m + 1$

(3) $\boldsymbol{\sigma}^{m} = \boldsymbol{\sigma}^{m} + (\boldsymbol{\sigma}^{m+1} - \boldsymbol{\sigma}^{m})/(\boldsymbol{m}^{\gamma} + \boldsymbol{N})$

(4) Repeat 2 and 3 until $\delta < \varepsilon$

(5) Calculate $q_{ip}^{m+1} = \dfrac{L_p\left(\boldsymbol{d}_i | \boldsymbol{H}_i, \boldsymbol{Z}, \boldsymbol{\sigma}_i; \Theta, \pi_p\right)}{L\left(\boldsymbol{d}_i | \boldsymbol{H}_i, \boldsymbol{Z}, \boldsymbol{\sigma}_i; \Theta, \pi\right)}$

(6) Calculate $\pi_p^{m+1} = \frac{1}{N}\sum_{i=1}^{N} q_{ip}^{m+1}$

(7) Calculate structural parameters $\boldsymbol{\theta}$ that maximizes

$$\boldsymbol{\theta}^{m+1} = \arg\max_{\theta} \sum_{i=1}^{N}\sum_{p=1}^{P}\sum_{t=1}^{T} \log\left[l\left(\boldsymbol{d}_{it} | \boldsymbol{H}_{it}, \boldsymbol{Z}, \boldsymbol{\sigma}_i; \Theta\right)\right]$$

(8) Go back to step 1 until the iteration converges

Intuitively, steps 1–4 explicitly solve the OE given structural parameters and the distribution of unobserved heterogeneity. Based on the estimated OE strategies in each iteration, we conduct the EM algorithm following Arcidiacono and Miller (2012) in steps 5–8 iteratively. The intuition behind the algorithm proposed by Arcidiacono and Miller is that we infer user types based on their observed decisions. For instance, when user A answers significantly more questions compared with user B with everything else equal, then user A is more likely to be in the segment with lower cost of answering questions.

We use recorded forum activity in the first 13 months as training period for customer support forum. Based on individual history of asking and answering questions, we calculate individual knowledge and social status level until the end of the training period. We use this knowledge and social status level as the initial value for our model estimation.

Identification

The variations in knowledge seeking and sharing behavior with respect to the state of their own knowledge and that of the other people in the community allow us to identify the effect of knowledge level and social status level. First of all, the updating rules for knowledge and reputation are different given an individual's actions; thus, the variation across the knowledge and reputation level can help us identify the effect of knowledge and reputation. An individual's knowledge-seeking decision will increase her knowledge level but decrease her social status level because her peers' social status increases. An individual's knowledge-sharing decisions will increase her social status while her knowledge level remains the same. Second, the knowledge seeking and sharing decisions critically depend on the knowledge increments from other people sharing their knowledge, which depend on the state of other individuals. If peers answer questions, the focal person's knowledge level increases, and her reputation decreases. This also helps identify the model.

Furthermore, the quality of answers generates exclusion restrictions between individual knowledge and social status. Because quality of an answer is assessed by all members of the community, the author of this answer has no control over the quality generation process. As a result, when an answer is selected as a high quality answer, social status of the author who posted this answer will be improved while her knowledge level remains the same. In the meanwhile, knowledge level of all other members increases, while their social status level remains the same.

Third, for individual knowledge seeking decision, our model setup also distinguishes the impact from social status from the impact from heterogeneous knowledge seeking cost. An individual asks a question based on the expected utility which is composed of expected increase in knowledge and the cost of asking a question, etc. The expected knowledge level increase is affected by the number of solutions provided, thus the status of the knowledge seeker and sharer, whereas the cost of asking only depends on the type of the person asking the question, and will not change over time irrespective of the social status of the knowledge seeker and sharer.

Predict View of Answers

For the customer support forum, we only have the aggregate number of views for each post, instead of individual viewing history for every post. As a result, we conduct a simulation where we randomly select individuals for viewership of each post based on their individual characteristics. In particular, we collect individual frequency of visiting the website by keeping track of the time when they last logged into the website on a daily basis for each one of the users in our dataset for a continuum of 3 months. This log-in frequency information, together with other individual characteristics, will influence the probability that individuals will be selected in the simulation.

We assume the probability that individual i views the answer posted by j for a question by o at period t as a logit function of the frequency of individual visiting the website, number of questions and answers individual i posted at period y:

$$\Pr\left(View_{i,s_{jot}}=1\right)=\frac{\exp\left(\varphi_0+\varphi_1 a_{it}+\varphi_2\sum_l s_{ilt}+\varphi_3 a_{it}*freq_i+\varphi_4\sum_l s_{ilt}*freq_i\right)}{1+\exp\left(\varphi_0+\varphi_1 a_{it}+\varphi_2\sum_l s_{ilt}+\varphi_3 a_{it}*freq_i+\varphi_4\sum_l s_{ilt}*freq_i\right)}$$

The probability of observing m views out of M individuals can be approximated using Poisson Binomial distribution. Thus we can use maximum likelihood estimation to estimate the parameters above. Given estimated parameters, we conduct simulations and randomly select individuals for viewership based on estimated probability of viewing each post.[1] Intuitively, this simulation allows us to randomly assign views among users weighted by their visiting frequency as well as asking and answering question decisions. We also set the upper limit of number of views for a user to 70 to prevent the possibility that active users view too many posts in one period.

The values of discount factors are chosen so that the community is in steady state (e.g., knowledge level, social status level, and the equilibrium policy). Our exploration of the dataset suggests that knowledge level is steady during our calibration period when we set the discount factor of knowledge level to be between 0.97 and 0.995. The steadiness of social status is a lot less sensitive to the discount factor compared with that for knowledge level. We also estimate our model with different values of β, and our estimation results still hold.

Appendix B

The Measure of the Degree of Core–Periphery Structure of a Network

A network has a core–periphery structure if individuals within the network can be divided into two parts: the core and the periphery. Individuals who belong to the core form links with others in the core and those who are in the periphery. However, those in the periphery do not form links with each other. For example, the network in Figure B1 is an ideal core–periphery network.

Figure B1 shows an example of a core–periphery structure with six users. The arrows represent answering questions. We can see from this figure that A, B, and C belong to the core group, in which they have connections with everyone. D, E, and F belong to the periphery group, in which they only have connections with core individuals. We use the following adjacency matrix to represent this ideal core–periphery network:

$$\mathrm{P}=\left[\begin{array}{ccc|ccc} & 1 & 1 & 1 & 1 & 1\\ 1 & & 1 & 1 & 1 & 1\\ 1 & 1 & & 1 & 1 & 1\\ \hline 1 & 1 & 1 & & 0 & 0\\ 1 & 1 & 1 & 0 & & 0\\ 1 & 1 & 1 & 0 & 0 & \end{array}\right]$$

Network A has six individuals; the first three individuals form the core of this network, and the remaining six individuals form the periphery of this network. This is an ideal core–periphery network because peripheral individuals do not have connections with each other, whereas the

[1] If we were given detailed information on viewing history for each post, we could further model the process in which individuals reduce the uncertainty on the number of views their answer will receive through observing the number of views answers on the forum received in the past.

only connections within this network are between core people or between core people and periphery people. The ideal matrix is called a *pattern matrix*. In this example, we show a symmetric network where elements in the matrix are either zero or one. However, the method can also be applied to our network, where the network is asymmetric, and the elements in the matrix can be larger than one.

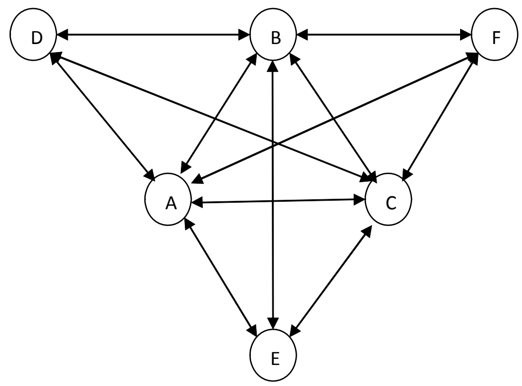

Figure B1. An Example of a Core–Periphery Structure

Borgatti and Everett (2000) propose a method for measuring the degree of core–periphery structure in a network by computing the Pearson correlation coefficient between the pattern matrix with the actual matrix: *Corr*(*A, P*), where *A* is the real matrix, and *P* is the pattern matrix. The higher the correlation, the more the network has a core–periphery structure. For an ideal core–periphery network, the coefficient is 1. For a matrix such as

$$P = \begin{bmatrix} & 1 & 1 & 1 & 1 & 0 \\ 1 & & 1 & 1 & 0 & 1 \\ 1 & 1 & & 0 & 1 & 1 \\ 1 & 1 & 0 & & 0 & 0 \\ 1 & 0 & 1 & 0 & & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}$$

the correlation coefficient is 0.7071. Below is a matrix that has no core group at all:

$$P = \begin{bmatrix} & 1 & 1 & 1 & 1 & 0 \\ 1 & & 1 & 1 & 0 & 1 \\ 1 & 1 & & 0 & 1 & 1 \\ 1 & 1 & 0 & & 0 & 0 \\ 1 & 0 & 1 & 0 & & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}$$

References

Arcidiacono, P., and Miller, R. 2011. "Conditional Choice Probability Estimation of Dynamic Discrete Choice Models with Unobserved Heterogeneity," *Econometrica* (79:6), pp. 1823-1867.

Borgatti, S., and Everett, N. 2000. "Models of Core–Periphery Structures," *Social Networks* (21:4), pp. 375-395.

Weintraub, G., Benkard, L., and Van Roy, B. 2008. "Markov Perfect Industry Dynamics with Many Firms," *Econometrica* (76:6), pp. 1375-1411.

Weintraub, G., Benkard, L., and Van Roy, B. 2010. "Computational Methods for Oblivious Equilibrium," *Operations Research* (58:4), pp. 1247-1265.