A Structural Model of Employee Behavioral Dynamics in Enterprise Social Media

APPENDIX

External Validity for Readership as a measure of reputation or expertise

We collected additional data on blog citations to provide external validity for our measure. Towards this, we acquired access to data on blog citations from the company that provided us the original data. The citations or hyperlinks on the Internet have been used as a measure of quality and influence as indicated by their use in search engine design (Brin and Page 1998). As is well known, these links are used to identify the most authoritative webpage for a given query. In fact the blog search engine, Technoratti, gives higher weights to posts with large number of links in their results. If readership measures reputation or expertise of a blogger then we should find posts by bloggers with higher readership levels, to have a higher probability of being cited by others.

In our dataset there are about 600 citations. We performed a logit regression to predict the probability that a post will be cited by others as a function of an individual's readership and other variables. The results of the key variables are reported in Table below:

Table 1: Logit regression predicting whether a work post will be cited or not.

Variable	Coefficient (standard error)
Constant	-9.894*** (1.014)
$log\ R_{itw}$	1.293*** (0.224)
K_{itw}	0.312** (0.157)
Status=High	0.078 (0.047)

Notes: Variables controlled for but not reported here include peer readership, peer knowledge, time dummies, topic dummies, individual leisure readership, and individual leisure knowledge. *** and ** denote significance at 1 and 5 percent, respectively.

Table 2: Logit regression predicting whether a leisure post will be cited or not.

Variable	Coefficient (standard error)
Constant	-8.741*** (2.795)
$log\ R_{itl}$	0.939*** (0.290)
K_{itl}	0.049* (0.027)

Status=High	-0.054 (0.079)

Notes: Variables controlled for but not reported here include peer readership, peer knowledge, time dummies, topic dummies, individual work readership, and individual work knowledge. *** and * denote significance at 1 and 10 percent, respectively.

These results show that posts of individuals with high readership are considered as being more influential by others and hence more likely to be cited by others. This provides some external validity that readership is a measure of reputation or expertise in blogosphere.

To provide additional validity for our measure, we used the help of two external coders who were experts on software testing to classify posts as high or low quality. We only focus on software testing category as that corresponds with the external coders' expertise. We provided these coders with all posts (447) of 119 individuals posted in Software Testing Category. We provided the coders with only the text of the post. The coders disagreed on only 6 posts. We discard these 6 posts for further analysis. There are 59 posts that were classified as high quality by these coders.

The basic idea about our readership measure providing utility to bloggers is that bloggers with higher readership would produce high quality posts and the audience will be able to identify them as experts on the topic. We find that the correlation between quality and readership is 0.91. This indicates that readers are attracted towards high quality posts. Further, we test the likelihood of bloggers coming up with high quality posts as a function of their readership.

Table 3: Logit regression predicting whether a software testing post is high quality or not (observation level is a post)

Variable	**Coefficient (standard error)**
Constant	-3.284*** (0.091)
$log\ R_{itw}$	0.578*** (0.149)
$log\ R_{itl}$	0.092 (0.174)
K_{itw}	0.881** (0.404)
K_{itl}	0.011 (0.083)
Status=High	0.091 (0.054)

Notes: *** denote significance at 1 percent.** denotes significance at 5 percent.

The results indicate that the individuals with high readership are more likely to produce high quality posts. This combined with the earlier result that high quality posts receive higher levels of readership imply that readers identify the bloggers who receive higher readership as the ones who produce high quality posts. Such identification by the readers would correspond to a higher reputation for bloggers in the enterprise setting.

External Validity for Knowledge

An external validity for our knowledge measure could be that knowledge of an employee from reading blogs as constructed by us leads to a higher performance rating for the employee. We have access to performance data on a subset of employees (272) in our sample. The performance data corresponds to an employee's performance from July 2007 to June 2008. The employee is evaluated by her immediate superior and rated on a discrete scale from 1 to 4 with 1 being the lowest and 4 being the highest rating. The ratings were provided at the end of June 2008. Because our blogging data corresponds to July 2007 to Oct 2008, it overlaps very well with the performance evaluation time period.

One of our arguments in the paper is that knowledge gained from reading blogs increases an employee's job performance. We use the data to test this relationship. We construct a measure of average work knowledge for an individual as the average of her knowledge from Nov 2007 to June 2008. (Recall that data from July 2007 to Oct 2007 is used as the hold-out sample for initial state calculation.) We first segment the individuals based on their job performance rating. For each rating, we calculate the average of the average knowledge of all individuals who received that rating. Below we show this graphically. As is obvious from the figure, individuals who are in a higher knowledge state on average receive a higher job performance evaluation.

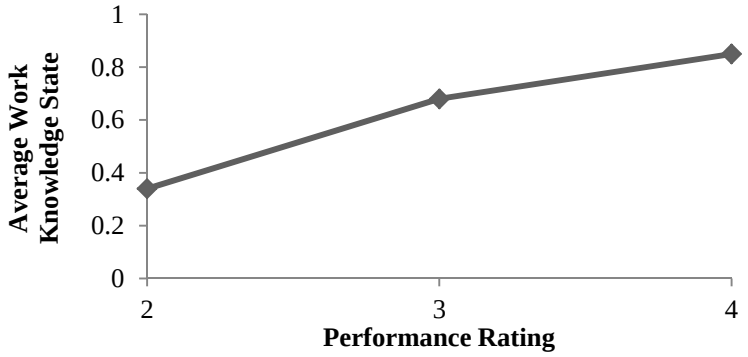

Figure 1: Knowledge State versus Performance Rating of Employees

We also constructed an ordered logistic regression to further establish the relationship between knowledge of an employee from reading blogs and her subsequent performance evaluation.

Table 4: Ordered Logistic Regression predicting Performance Rating

Dependent Variable: Performance Rating (observation level is employee)

Variable	Parameter Estimate	Standard Deviation
K_{iTw}	2.17*	1.15

$\sum_{t=1}^{T} K_{itw}/T$	7.94***	1.25
Cut Point 1	1.65***	0.45
Cut Point 2	7.47***	0.75

Notes: Variables controlled for but not reported here include individual average readerships on work and leisure, peer average readerships on work and leisure, peer average knowledge of work and leisure. *** and * denote significance at 1 and 10 percent, respectively.

The results indicate that individuals who are on average more knowledgeable (as constructed by us) receive a higher performance evaluation. We acknowledge that we only have access to a cross section of performance data, and hence, we cannot conclude a causal relationship between knowledge gained from blogging and job performance in the workplace. However, we can at least see that the measure of knowledge as constructed by us for our analyses is strongly associated with job performance.

ESTIMATION PROCEDURE

Step 1. Estimate the transition probabilities for $\log(R_{itw})$, $\log(R_{itl})$, K_{itw}, and K_{itl} and the Conditional Choice Probabilities (*CCP*).

All the states have discrete ordered levels. As a result, we model the state transition as an ordered logistic regression. We allow the regression parameters to be current state specific. The CCP represents the probabilities of choosing actions given the state values. Given the states, we can estimate the CCP through a multinomial logistic regression of actions on state variables. The *CCP* and the state transitions are jointly estimated through a *HMM*.

The *HMM* operates as follows: The state transition probabilities for $\log(R_{itw})$ and $\log(R_{itl})$ are a function of the variables given in Equation 9. Further, the state transitions for K_{itw}, and K_{itl} are functions of the variables given in Equation 10. In Equations 9 and 10, we employ the discretized values of cumulative readerships when calculating the readership specific variables. Let $D_i = \{a_{i1}, a_{i2}, \ldots., a_{iT}\}$ represent the sequence of choices for individual *I*, S_i=$\{s_{i1}, s_{i2}, \ldots., s_{iT}\}$ represent the state sequence over time for individual *i*, π_i represent the initial state distribution (*1xNS*) for individual *i*, and λ represent the set of parameters that govern the state transition probabilities and the *CCP (NSx64*). An element *jk* of *CCP* represents the probability of action *k* given state *j*. For simplicity, let us represent the observed state by *O* and the unobserved/hidden state by *H*. Further, let O_i and H_i represent the observed and unobserved state sequences for individual *i*. Thus, the probability of the observed outcome sequence D_i and observed state sequence O_i is obtained (Rabiner 1989) as follows

$$L(D_i, O_i) = P(D_i, O_i|\lambda) = \sum_{\forall H_i}\big(P(D_i|\lambda, O_i, H_i) * P(O_i|\lambda, H_i) * P(H_i|\lambda)\big). \quad (18)$$

We maximize Equation 18 to obtain parameter set λ. The initial state distribution is analytically derived by solving the equation $\pi_i = \pi_i \bar{Q}$, where $\bar{Q}$ is the state transition matrix calculated at the mean value of covariates for individual *i*. The number of levels in the knowledge states is determined by comparing AIC from models with different numbers of levels in the knowledge states. Note that the individual state transitions in Step 1 are function of peers' states of which the knowledge state is unobserved. To address this issue, we identify the probability for individual k being in unobserved state s at time t. We use these probabilities to calculate the parameters for the individual state transition matrices.

Step 2. Calculate Utility Parameters

In this step, we estimate the structural parameters: $\rho = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8, \theta_9, \theta_{10}, \theta_{11}, \theta_{12})$. In an OE, the *CCP* is a function of an individual's own state. Peers influence the utility of an individual because the *CCP* and the state transition probabilities determine the steady state peer state distribution[1]. The main computational advantage that OE provides is to convert the multi-agent dynamic game problem into a single agent dynamic optimization problem. Moreover, one can use any one of several methods proposed in the literature to solve this single agent dynamic optimization problem. We follow Agurregabiria and Mira (2007)'s nested pseudo maximum likelihood procedure to solve the single agent dynamic optimization problem.

The estimation procedure operates as follows.

Iterate through the following steps until $\rho^m \sim \rho^{m+1}$, where ρ^m is the value of the structural parameters in the m^{th} iteration.

1. Given a steady state peer state distribution, parameter set λ, and *CCP*, calculate action-specific state transition matrices (Q_a). The peer-specific variables required to calculate the state transition probability are calculated using the steady state peer state distribution and *CCP*. This is possible because peers also follow the same *CCP*.

2. Using the *CCP* and the action-specific state transition matrices, calculate the unconditional state transition matrix (Q) as $Q = \sum_{a=1}^{64} CCP(.,a) * Q_a$. Here $CCP(.,a)$ is the column *a* of *CCP*.

3. Calculate the state-specific value function V as

$$V = (I - \beta Q)^{-1} * \left[\sum_{\forall a} CCP(.,a) \odot (u(a) + \gamma(a))\right]$$

[1] The peer state distribution indicates the number of peers in each state.

Here *I* is an identity *(NSxNS)* matrix, $u(a)$ is an *NSXnp* action-specific expected profit matrix before the random shock is realized (*np* is the number of structural parameters to be estimated), $\gamma(a)$ is an *NSx1* action-specific random shock vector, and $\odot$ represents an element by element multiplication. Note that $u(a)$ is the current period utility absent the random shock.

4. Calculate the vector of action-specific value function, V_a as

$$V_a = u(a) + \beta * Q_a * V.$$

5. Find structural parameter set ρ by maximizing $L(D_i, O_i)$

-Given a value of ρ and V_a, calculate $P(a|s)$ as

$$P(a|s) = \frac{exp(V_a(s))}{\sum_{\forall a} exp(V_a(s))}.$$

Here, $V_a(s)$ is the expected value of taking action *a* for an individual in state *s*.

-Use $P(a|s)$, calculated above, λ and π_i, calculated in Step 1, to calculate $L(D_i, O_i)$.

- Find ρ by maximizing $L(D_i, O_i)$.

6. Using ρ and V_a, update the *CCP*.

7. Calculate the steady state peer state distribution.

-update Q using Q_a and *CCP*.

-calculate *n* period state transition probability matrix Q^n (Weintraub et al. 2008); n should be sufficiently large such that $Q^n \approx Q^{n-1}$

-the long-run peer state distribution is any row of Q^n.

Note that in Step 1, the calculation of $L(D_i, O_i)$ is conditional on the steady state peer state distribution. In essence, Step 1 is an inner loop for Step 2.