Seeking Variety: A Dynamic Model of Employee Blog Reading Behavior

Param Vir Singh, Nachiketa Sahoo, Tridas Mukhopadhyay*
{ psidhu, nsahoo, tridas }@cmu.edu

Tepper School of Business
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213

MAY 2010

*Param Vir Singh is Assistant Professor of Information Systems, Nachiketa Sahoo is Assistant Professor of Information Systems, and Tridas Mukhopadhyay is Deloitte Consulting Professor of e-Business, all at the Tepper School of Business at Carnegie Mellon University. Authors thank ILAB at the Hienz College, Carnegie Mellon University for their generous support and data that made this study possible. Contact address: tridas@cmu.edu.

Seeking Variety: A Dynamic Model of Employee Blog Reading Behavior

Abstract

We investigate the dynamics of blog reading behavior of employees in an enterprise blogosphere. A dynamic model is developed and calibrated using longitudinal data from a Fortune 1000 IT services firm. We identify a *variety-seeking* behavior of blog readers where they frequently switch from reading on one set of topics to another dynamically. Our results indicate that this switching behavior is induced by the textual characteristics (*sentiment* and *quality*) of the posts read, reader characteristics (*status*, *location*, *expertise*), or a readers' inherent desire for variety. Our modeling framework allows us to segregate the impact of post-textual characteristics on attracting readers from retaining them. We find that the textual characteristics that appeal to the *sentiment* of the reader affect both reader attraction and retention. However, textual characteristics that reflect only the *quality* of the posts affect only reader retention. The modeling framework and findings of this study highlight opportunities for a firm to influence blog reading behavior of its employees to align it with its goals. We provide directions to improve the utility of blogs as a medium for knowledge sharing. Overall, the blog reading dynamics estimation of this study contributes to the development of theoretically grounded understanding of reading behavior of individuals in online settings and more specifically in communities formed around user generated content.

Keywords: Blogs, Employee Blogs, Enterprise 2.0, Blog Reading, Dynamic Models, User Generated Content, Latent Class Models, Hidden Markov Models, Regime Switching Models, Text mining.

1 Introduction

Blogs are one of the earliest Web 2.0 technologies to be adopted within organizations. Prominent adopters of organizational blogs are Microsoft, Sun Microsystems, Google, Adobe, IBM, Infosys, etc. Employees spend considerable time in designing, thinking, reading, and writing posts. It has been argued that the value of knowledge shared among employees through blog posts makes up for such expense (Mowbray 2004, Efimova 2004, Huh et al 2008, Huang et al 2010).

Encouraged by positive media reports and given the lukewarm acceptance of traditional knowledge sharing initiatives, firms are increasingly adopting non-traditional channels such as blogs and wikis for sharing knowledge among employees (Huang et al 2010, Lucier and Torsilieri 1999, Efimova 2004, Huh et al 2008). Examining blogs at a large IT services firm, we found that fewer work relevant articles are being *published* than non-work relevant articles. Moreover, work relevant articles are *read* less often, when compared on a per post basis, than non-work relevant posts. For every work relevant article, two non-work relevant articles are published. Further, for every one person reading a work relevant post four people read non-work relevant posts. This raises the question whether work related knowledge is being assimilated through blogs.

Firms are confronted with the challenge of promoting posting as well as reading of content that helps employees in their work. To facilitate effective sharing of knowledge a firm needs to understand the reading behavior of its employees so that they can be directed to better align with the firm's objectives. Indeed, understanding the reading behavior is crucial to promote posting because, by only understanding the reading behavior, a firm can guide bloggers about the type of content that would be well accepted by the readers. However, research in understanding blog reading behavior has been difficult since *blog reading*, unlike publishing on blogs or commenting on blogs, is usually invisible to external observers. In this study we take a step toward shedding some light on the blog reading behavior of employees with the

help of a unique dataset collected from a large IT services firm. We seek to understand this behavior and how it changes so that an organization may take steps to influence it to its advantage.

The objective of this research is to capture the dynamics of employee blog readership. In this exploratory study, we set out to examine what characteristics of the blog posts, bloggers or readers determine blog reading behavior. Our model allows us to probabilistically identify the reader's latent state at any given time and enables us to understand how her encounters with bloggers and post characteristics trigger her to move to a different state of readership. Using our modeling framework, we are able to ask a series of interesting questions: How does the readability, comprehensiveness, negativity, controversialness or grammatical errors in posts affect the magnitude of readership and the probability of the reader switching from one state to another? What role does the location (e.g., U.S. vs. non-U.S.), specialization (technical vs. non-technical), or organizational status (high vs. low) of the employee play in her reading behavior?

We model the readership dynamics in the same spirit as the dynamic state space models by Hamilton (1989) and Rabiner (1989). There are two key advantages of our dynamic state space model over the traditional static models when studying individual behavior. The first advantage is that a static model may give erroneous estimates of factors affecting behavior if there are dynamics in the behavior. Our model allows for dynamics in individual behavior without enforcing them. Hence, when there are dynamics in the individual behavior, our model will be superior to static models and provide consistent estimates of factors that affect behavior. In contrast, in the absence of dynamics in behavior, the dynamic state space model would collapse to a static model. The second advantage of the dynamic model is its ability to provide insights into when and how the individual behavior should be influenced. More specifically, it allows the segmentation of individuals according to their reading behavior in terms of types and volume of posts read at any time. This segmentation allows one to investigate the varying impacts of blogger and post characteristics across individuals on their movement from one reading

behavior to another. This provides insight into the strategic targeting of individuals in different states to effectively induce higher work related readership. Such insight cannot be obtained in the standard static framework.

The main contribution of this research is in providing a behaviorally structured model that provides managers with the ability to influence blog reading behavior of employees. Using our model the manager can dynamically classify an employee to a reading state, and assess the dynamic effects of post, and reader characteristics on employee transitions to the most preferable state. The other key contributions of this study pertain to the results obtained using our dynamic model. We find a *variety-seeking* behavior displayed by blog readers whereby readers frequently switch from reading on one set of topics to another. These switching are found to be induced by post (both *quality* and *sentiment*) and reader characteristics (*organizational status, location, expertise*) as well as an inherent desire for variety. Finally, we find that readers rarely switch directly from reading intensely on one set of topics to another and this transition is interspersed by an exploratory phase where they sample from different topics.

1.1 Research Setting

Our research setting is a large, Fortune 1000 information technology services, business process outsourcing, and consulting firm. Fortune named this firm one of the fastest growing companies in 2009. Its annual revenues in year 2009 were a few billion dollars and it employs more than 75,000 employees. It is a United States based firm with significant operations in other countries in Europe, Asia, and Americas. Over the years this led to the emergence of localized silos of knowledge and little knowledge flow from one location to another. Even within the same location, the knowledge sharing was not efficient. To facilitate knowledge and information sharing across as well as within locations, the firm has undertaken several measures. Prominent among these measures is the use of Web 2.0 technologies within the enterprise.

This firm has been one of the earliest adopters of Web 2.0 technologies in its industry. Given a primarily technologically educated workforce, it is not surprising that these technologies have been well accepted by its employees. The focus of this study is on employee blogs for this firm.

This firm adopted the use of employee blogs in mid 2006. These blogs are hosted on an internal platform and are accessible only to the employees of the firm. Every employee is allowed to maintain her own blog on this platform. This blog is accessible to all the employees of the firm. The identity of the blog writer is also revealed on the blog. Bloggers classify their posts into one of the 25 categories (for example, software testing, movies, history, knowledge management, senior management, etc). The firm does not restrict the kind of posts that can be written by the employees.

To analyze the type of content that is being shared on the internal blogosphere, the firm broadly classifies the blog article categories into four broad topics. These four topics are: Non-Work Entertainment Posts (NWE), Non-Work Knowledge Posts (NWK), Work Related Managerial Knowledge Posts (WMK), and Work Related Technical Knowledge Posts (WTK). Table 1 presents the sub categories that constitute each topic.

Table 1. Blog Post Classification

Topics	NWE	NWK	WMK	WTK
Sub Categories	Fun; Movies-TV-Music; Sports; Puzzles; Chip-n-putt	Religion-Spiritual-Culture; History-Culture; Education-Motivation; Photography; Arts; Poetry-Stories; Corporate Social Responsibility; Books; Geographies	Corporate Functions; Knowledge Management; Project Management; Business Development; Senior Management; Feedback; Practices-Programs-Accounts	Free/Libre/Open Source Software; Technology; Testing; Domains

Approximately two out of three posts have been in the Non Work topics (NWE and NWK). Similarly, there has been higher readings in Non-work related categories (1.2 million) compared to work

related categories (0.24 million). The writing and reading of blogs is voluntary and the firm does not provide any direct incentives for using these Web 2.0 technologies. While the upper management of the firm wants employees to read more work related topics, they do not want to constrain employee behavior in any way that may reduce the employees' use of the blogging platform.

1.2 Literature Review

This study relates to the emerging literature on Blogs. Several aspects of blogs have been studied by researchers over the years. These studies have contributed to our understanding by systematically analyzing how blogs affect products sales (Goldstone 2006, Mishne and Glance 2006), outcome of important events (Adamic and Glance 2005, Farrell and Drezner 2004), new venture funding (Aggarwal et al 2010), perceptions about the firm (Aggarwal et al 2009), and culture (Blood and Rodzvilla 2007). They have also shed light on why people blog (Nardi et al 2005, Miura and Yamashita 2007, Furukawa et al 2007), why bloggers cite others' posts, and leave comments on posts (Adamic and Glance 2005), and how information diffuses in blogosphere (Adar and Adamic 2005, Leskovec et al 2007). Several researches have also developed ways to detect communities in blogosphere (Lin et al 2006) and discover important bloggers (Nakajima et al 2005).

Recently, there has been an impetus to study blog reading behavior (Furukawa et al 2007, Adamic and Glance 2005). These studies explain why individuals cite and comment on other posts, and join blog-rolls. They focus on readers who are also bloggers. They attempt to infer blog reading from observable activity on the blogs. However, to our knowledge, there is no study that systematically analyzes the dynamics in blog reading behavior using actual observations of blog reading. A primary reason for such a gap in research is the unavailability of blog reading data over time and detailed information about readers at large. However, such data are available to the blog hosts in the form of web server access logs. This research has been made possible due to a successful sharing agreement of such data between the host company and our research team.

This study also relates to the literature on broader user generated content. While there are several studies which have shown the impact of user generated content on consumer choices (e.g. Chevalier and Mayzlin 2006), recent studies have started to shed light on what aspects of the content have the most effect. Ghose and Ipeiroitis (2010) find that readers find grammatically correct reviews with moderate objectivity and readability to be most helpful. Lu et al (2010) also find similar results in the context of Epinions.com and further emphasize the role of review's recency as an important attribute that determines its usefulness. Forman et al (2009) find that the reviewer identity plays a key role in whether the readers trust her review or not. In the context of blogs, Aggarwal et al (2009) find that readers find blogs with a mixture of positive and negative posts more credible. Finally, Huang et al (2010) study the different incentives and tradeoffs involved in consuming and contributing user generated content.

This research also draws from the literature on behavior dynamics. Research in the area of behavior dynamics is emerging both on the theoretical (e.g. McAlister and Pessemier 1982, McAlister 1982) and empirical methodological perspectives (e.g. Hamilton 1989, Montgomery et al 2005, Singh et al 2010, Netzer et al 2007, Bolton and Lemon 1999, Li et al 2002). From the theoretical perspective the focus has been on understanding the drivers behind the dynamics. The behavioral dynamics have been attributed to variations in situational variables (e.g., individual's environment or interactions) (Aaker, Fournier and Brasel 2004, Fournier 1998) and psychological factors (curiosity and boredom) (McAlister and Pessemier 1982, McAlister 1982). Oliver (1997 and 1999) suggests that situational variables can increase or decrease satisfaction of an individual and when they increase / decrease beyond a certain threshold the individual is likely to switch to a conceptually different behavior. On a similar note, situational factors can affect the extent of satisfaction that a reader can derive from reading on a specific topic and hence can affect the readership dynamics. On the psychological side, due to boredom, the utility from one item decreases with repeated use while competing items become more attractive simply by being not the current item. Curiosity implies that different items may hold promise of providing different experiences which may cause an individual to try new items. It is obvious that both these factors are

relevant to the case of blog reading. Reading on a topic repeatedly can lead to boredom and hence the reader may switch away from the topic. At the same time new topics may provide different experience and hence they would induce individuals to try them. Hence, it is likely that employees will exhibit dynamics in their reading behavior where they would switch from one topic to another. However, the extent to which these dynamics would be present is not obvious.

The empirical methodological literature in this area provides some guidance for modeling the dynamics in behavior. Heckman (1984) suggests modeling the behavior dynamics as *state dependence*. A number of studies have followed this approach. However, except for a few, most studies have incorporated *state dependence* through a single variable (e.g. number of past purchases) that accounts for the past in an otherwise static model (e.g. Guadagni and Little 1983, Srinivasan and Kesavan 1976). Recent studies by Singh et al (2010) and Netzer et al (2007) have emphasized that *state dependence* should be modeled more comprehensively. They argue that a single state variable in an otherwise static model may not be able to capture the true dynamics. Besides, it does not allow us to structurally account for the impact of situational variables which may be of interest to managers. In this study, we account for dynamics in a more elaborate fashion by allowing the parameters of an individual to change dynamically as a result of situational variables and unobserved effects.

The rest of the paper proceeds as follows. In Section 2, we develop three models of blog reading behavior – a simple model, latent state model, and dynamic model. In Section 3, we describe our dataset along with variable definitions. In Section 4, we present the results from the dynamic model and in Section 5 we present post hoc analyses, key findings, managerial implications and the contributions of this study. In Section 6, we conclude by discussing the limitations of our study and guidelines for potential future research.

2 Modeling Framework

The modeling framework that we employ consists of three parts: (1) a simple model of blog reading, (2) a latent state model of blog reading, and (3) a dynamic model of blog reading. All three models are designed to explain the employees' reading volume on a topic in a month. This was chosen because one objective of the firm is to have the employee read more on the topic that is of interest to the firm, e.g., posts with technical knowledge, or posts with managerial knowledge.

In the simple model of blog reading each blog reader's amount of reading in a topic is modeled in the same functional form with a common set of parameters. This reading behavior does not change over time. A latent state model of blog reading implies that readers can be divided into several latent states based on their reading behavior (Greene and Hensher 2003). However, the reading behavior of an individual is not allowed to change over time. A dynamic model of blog reading allows readers to belong to different latent states of reading behavior over time (Hamilton 1989, Rabiner 1989). Hence, it allows readers to move from one state to another over time. It is easy to see that a simple model is a special case of a latent state model which is a special case of a dynamic model. In essence, the dynamic model of blog reading developed here would incorporate cross sectional (across readers) as well as inter temporal (across time for the same reader) unobserved heterogeneity in behavior. The latent state model would capture only the cross sectional unobserved heterogeneity and the simple model would incorporate none. Note that both the simple as well as latent state models are static models of behavior.

2.1 A Simple Model of Blog Reading

Let N_{ijt} be the count of posts read by individual i on Topic j in time period t . Let $N_{it} = \{N_{i1t}, N_{i2t}, N_{i3t}, N_{i4t}\}$ represent the set of readings across topics for individual i at time t . It is likely that the counts across topics for an individual are jointly determined. Since there are four topics, we model their counts as a Quad-variate Negative Binomial Model. The joint distribution of the counts can be given as:

$$h(N_{i1t}, N_{i2t}, N_{i3t}, N_{i4t} | \alpha, \lambda_{1t}, \lambda_{2t}, \lambda_{3t}, \lambda_{4t}) = \int_0^\infty \left\{ \prod_j \left[(\lambda_{jt} \theta)^{N_{ijt}} \frac{\exp(-\lambda_{jt} \theta)}{N_{ijt}!} \right] \right\} \times \left[(\theta)^{\alpha-1} \frac{\exp(-\theta)}{\Gamma(\alpha)} \right] d\theta,$$

which, following Arbous and Kerrich (1951), can be rewritten as:

$$h(N_{i1t}, N_{i2t}, N_{i3t}, N_{i4t} | \propto, \lambda_{1t}, \lambda_{2t}, \lambda_{3t}, \lambda_{4t}) = \frac{\Gamma(\alpha + \sum_j N_{ijt})}{\Gamma(\alpha) \prod_j N_{ijt}!} \left(\frac{1}{\eta} \right)^\alpha \prod_j \left(\frac{\lambda_{jt}}{\eta} \right)^{N_{ijt}} \quad (1)$$

with $\eta = 1 + \sum_j \lambda_j$. Here, λ_j is the parameter for N_{ij} , and following the standard practice in Econometrics literature is parameterized as $\lambda_j = \exp(z_j' \beta_j)$. z_j' is the vector of covariates that affect the average amount of reading by an individual on Topic j . It may consist of Topic related variables and other controls. Hence, the likelihood of the data under the simple model of reading behavior can be written as:

$$L(N | \propto, \beta_1, \beta_2, \beta_3, \beta_4) = \prod_i \prod_t h(N_{i1t}, N_{i2t}, N_{i3t}, N_{i4t} | \propto, \beta_1, \beta_2, \beta_3, \beta_4).$$

2.2 A Latent State Model of Blog Reading

Let us assume there are S number of latent states of Blog reading behavior. Further, let β_{js} be the set of parameters that capture the distinct behavior of individuals in state s . For behavior to be distinct across states the set of parameters β_{js} should vary across S states. To understand it better let us assume that z_j is a vector of ones. Then if $\beta_{js} > \beta_{js'}$ it would imply that individuals in latent state s read more on topic j on an average than the amount of topic j read on an average by individuals in state s' . Since we have a Quadrivariate Negative Binomial model, we do not restrict the variables in the β_{js} to be different across states. For example, individuals in states s and s' may read the same amount of topic j but differ on topic j' . The likelihood of observing N_{it} for an individual in state s can be written as:

$$L(N_i | \propto, \beta_{1s}, \beta_{2s}, \beta_{3s}, \beta_{4s}, i \in s) = h(N_{i1t}, N_{i2t}, N_{i3t}, N_{i4t} | \propto, \beta_{1s}, \beta_{2s}, \beta_{3s}, \beta_{4s}),$$

where, the function $h(N_{i1}, N_{i2}, N_{i3}, N_{i4} | \propto, \beta_{1s}, \beta_{2s}, \beta_{3s}, \beta_{4s})$ is as in Equation 1.

Let p_{is} be the probability that individual i belongs to state s . Then $\sum_s p_{is}=1$. The probability p_{is} can be construed to be a function of individual demographics and socio-economic status variables (D_i). The state assignment can be modeled as a multinomial Logit process such that:

$$p_{is} = \frac{\exp(\delta'_s D_i)}{1 + \sum_{s=1}^{S-1} \exp(\delta'_s D_i)},$$

where δ'_s is the set of parameters for state s that correspond to D_i .

The two equations can be combined to get the likelihood of N_{it} for individual i unconditional on her state as:

$$L(N_{it} | \alpha, \beta_{11}, \dots, \beta_{4S}) = \sum_s p_{is} \times h(N_{i1t}, N_{i2t}, N_{i3t}, N_{i4t} | \alpha, \beta_{1s}, \beta_{2s}, \beta_{3s}, \beta_{4s}).$$

And the likelihood of the entire dataset can be written as:

$$L(N | \alpha, \beta_{11}, \dots, \beta_{4S}) = \prod_i \prod_t \left[\sum_s p_{is} \times h(N_{i1t}, N_{i2t}, N_{i3t}, N_{i4t} | \alpha, \beta_{1s}, \beta_{2s}, \beta_{3s}, \beta_{4s}) \right].$$

Note that the state in which an individual belongs is determined endogenously and is not pre-specified by the Econometrician. The number of states S is also a priori unknown and is determined by comparing the Bayesian Information Criterion for models with different number of states.

2.3 A Dynamic Model of Blog Reading Behavior

A dynamic model allows both cross sectional and inter temporal heterogeneity. As in the case of Latent state model let us assume that there are S latent states of blog reading behavior, β_{js} be the set of parameters that capture the distinct behavior of individuals in state s , and some components of β_{js} vary across states to allow for distinct behavior across states. Unlike the latent state model, we now allow individuals to move across states over time. If the state sequence (C_i) of an individual (in which state the

individual is at each time) is known, one can write the likelihood function for her observed reading over time conditional on the observed state sequence as:

$$L(N_i|\alpha, \beta_{11}, \dots, \beta_{4S}, C_i) = \prod_t h(N_{i1t}, N_{i2t}, N_{i3t}, N_{i4t}|\alpha, \beta_{1s}, \beta_{2s}, \beta_{3s}, \beta_{4s}, C_{it} = s),$$

where the function $h(N_{i1t}, N_{i2t}, N_{i3t}, N_{i4t}|\alpha, \beta_{1s}, \beta_{2s}, \beta_{3s}, \beta_{4s}, C_{it} = s)$ is as in Equation 1.

The likelihood of observing the state sequence and the reading pattern jointly can be written as:

$$L(N_i, C_i|\alpha, \beta_{11}, \dots, \beta_{4S}) = L(N_i|\alpha, \beta_{11}, \dots, \beta_{4S}, C_i) \times P(C_i)$$

where $P(C_i)$ is the probability of observing this particular state sequence for individual i . $P(C_i)$ can be construed as a combination of two components: initial state distribution (π) and the state transition distribution (Φ). The initial state distribution is represented as an $S \times 1$ vector:

$$\pi = \begin{bmatrix} p(C_{i1} = 1) \\ p(C_{i1} = 2) \\ \vdots \\ p(C_{i1} = S-1) \\ p(C_{i1} = S) \end{bmatrix}.$$

The elements of the initial state distribution represent the probability that an individual is in the corresponding state in time period 1. For example, the first element of π is the probability that an individual is in state 1 in period 1.

The state transition distribution is represented as a $S \times S$ matrix:

$$\Phi = \begin{pmatrix} \varphi_{11} & \varphi_{12} & \dots & \varphi_{1S-1} & \varphi_{1S} \\ \varphi_{21} & \varphi_{22} & \dots & \varphi_{2S-1} & \varphi_{2S} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \varphi_{S-11} & \varphi_{S-12} & \dots & \varphi_{S-1S-1} & \varphi_{S-1S} \\ \varphi_{S1} & \varphi_{S2} & \dots & \varphi_{SS-1} & \varphi_{SS} \end{pmatrix},$$

where $\sum_c \varphi_{rc} = 1$ for $\forall r$. The element corresponding to row $r \in (1, \dots, S)$ and column $c \in (1, \dots, S)$, φ_{rc} represent the probability that an individual who is in state r in period t will switch to state c in period $t + 1$.

Given, the initial state distribution and the state transition matrix, we can write the likelihood of the observed reading pattern of an individual and the state sequence jointly as:

$$L(N_i, C_i | \alpha, \beta_{11}, \dots, \beta_{4S}) = \pi_{C_{i1}} \times L(N_{i1} | \alpha, \beta_{1C_{i1}}, \dots, \beta_{4C_{i1}}) \times \prod_2^T \varphi_{r=C_{it-1}, c=C_{it}} \times L(N_{it} | \alpha, \beta_{1C_{it}}, \dots, \beta_{4C_{it}}).$$

If we take the expectation of the above likelihood over all possible state sequences for an individual we can get the unconditional likelihood for the observed reading pattern for an individual i as:

$$L(N_i | \alpha, \beta_{11}, \dots, \beta_{4S}) = \sum_{C_i} \left[\pi_{C_{i1}} \times L(N_{i1} | \alpha, \beta_{1C_{i1}}, \dots, \beta_{4C_{i1}}) \times \prod_2^T [\varphi_{r=C_{it-1}, c=C_{it}} \times L(N_{it} | \alpha, \beta_{1C_{it}}, \dots, \beta_{4C_{it}})] \right].$$

The likelihood of the entire dataset can be written as:

$$L(N | \alpha, \beta_{11}, \dots, \beta_{4S}) = \prod_i \sum_{C_i} \left[\pi_{C_{i1}} \times L(N_{i1} | \alpha, \beta_{1C_{i1}}, \dots, \beta_{4C_{i1}}) \times \prod_2^T [\varphi_{r=C_{it-1}, c=C_{it}} \times L(N_{it} | \alpha, \beta_{1C_{it}}, \dots, \beta_{4C_{it}})] \right].$$

A few points need to be emphasized here. The state transition probability φ_{rc} assumes that the probability to transition to a state in time $t + 1$ is only the function of the state the individual is in time t . That is, the state transition follows a Markovian process. Several other studies have made similar Markovian transition assumption in their dynamic models (e.g. Hamilton 1989, Singh et al 2010, Netzer

et al 2007). This is a simplifying assumption. One could construe a more sophisticated state transition matrix where, for example, the probabilities depend on the entire history of state sequence of the individual. We have used the simple Markovian process for state transition to keep the model parsimonious and the number of parameters to be estimated manageable. The state dependency takes care of the serial correlation that may exist among the errors (Hamilton 1989).

The state transition matrix shown above is assumed to be constant over time. We can make the transition matrix vary over time by making the transition probabilities a function of variables that may vary over time. For example, one such variable could be the length of posts read by an individual in the current period. Let X_{it} be the vector of such variables that affect the state transition probabilities. The state transitions can also be affected by random factors unobservable to the researcher. We can model the state transition probabilities as coming from a multinomial Logit regression as:

$$\varphi_{rcit} = \frac{\exp(\rho_r' X_{it})}{1 + \sum_1^{S-1} \exp(\rho_r' X_{it})}, \quad \sum_c \varphi_{rcit} = 1.$$

Here, ρ_r is a vector of state specific parameters. Under this formulation the likelihood of the full dataset can be written as:

$$L(N|\alpha, \beta_{11}, \dots, \beta_{4S}) = \prod_i \sum_{C_i} \left[\pi_{C_{i1}} \times L(N_{i1}|\alpha, \beta_{1C_{i1}}, \dots, \beta_{4C_{i1}}) \times \prod_2^T [\varphi_{r=C_{it-1}c=C_{it}it} \times L(N_{it}|\alpha, \beta_{1C_{it}}, \dots, \beta_{4C_{it}})] \right]$$

Till now our dynamic model only incorporates inter-temporal heterogeneity. We further introduce cross sectional (individual level) heterogeneity in the dynamic model through a hierarchical Bayesian framework. To keep the number of parameters manageable, we allow the coefficients corresponding to the constant terms in the state transition parameters (ρ) and the reading parameters (β) to be individual specific. Hence, in our model, β_{0ijs} and ρ_{0is} are reader specific random effects parameters. Observed

individual characteristics that may affect the reading levels and state transitions are introduced into the model in a hierarchical manner (Allenby and Ginter 1995). Let $\theta_i = \{\beta_{0i11}, \dots, \beta_{0i1S}, \beta_{0i21}, \dots, \beta_{0i2S}, \beta_{0i31}, \dots, \beta_{0i3S}, \beta_{0i41}, \dots, \beta_{0i4S}, \rho_{0i1}, \dots, \rho_{0iS}\}$ be the set of random-effect parameters. It can be written as a function of observed and unobserved individual characteristics as:

$$\theta_i = \boldsymbol{\vartheta}' \mathbf{D}_i + \varepsilon_{\theta_i},$$

where $\mathbf{D}_i$ is a vector of individual characteristics for individual i , $\boldsymbol{\vartheta}$ is a matrix of parameters, relating the individual characteristics to the random effect parameters (θ_i), and $\varepsilon_{\theta_i} \sim N(0, \Sigma_{\theta})$. Figure 1 presents the three models graphically.

3 Data

Data for our analysis come from three archival sources of the firm:

1. The Blog Post Database contains complete details about blog posts. In this database, each blog and each post are uniquely identified. It also contains blogger ID, post content, posting date, blogger specified post classification (in one of the 25 categories shown in Table 1), and comments made on the post.
2. The blog reading database contains who read what post at what time. We have access to this information from *Jan 1 2007* to *Oct 31 2008*.
3. The Human Resource database contains information on employee demographics (Gender and Country), position in the organization, and employee department within the organization. Note that the firm identifies bloggers and readers by their unique employee IDs.

Figure 1 Graphical Representation of Models of Blog Reading

These three sources give us rich information over a period of 22 months. Approximately 15,000 employees have written about 71,000 posts during this period. On average each post was 300 words long. There have been approx 286,000 replies (on average 33 words long) posted by readers to these posts indicating a healthy user involvement.

We randomly selected 1,200 unique readers and split them into two datasets: 1000 readers in calibration dataset and 200 readers for test dataset. The calibration dataset has approx 74% males, 9% high status, 25% middle status, 59% Asians, 19% North Americans, and 67% technical readers. There are a total of approx 38,000 post readings by these 1000 employees during the 22 month period, 44% of which are in the NWE, 36% in NWK, 17% in WMK, and the remaining in WTK.

3.1 Variable Construction

We constructed three sets of variables based on our discussion in Section 2 (Table 2 A). The first set includes variables that affect the amount of posts read in each period per category. We call this the reading set. It corresponds to the matrix $\mathbf{Z}$ in Sections 2.1, 2.2, and 2.3. The second set includes variables that affect the state transition probabilities. We call this the state transition set. The state transition set is only relevant to Section 2.3 and corresponds to matrix $\mathbf{X}$. The third set of variables constitutes reader characteristics that are used to account for reader specific heterogeneities for both the amount of reading and state transitions. We call this the reader characteristics set. This corresponds to matrix $\mathbf{D}$ in Sections 2.2 and 2.3.

3.1.1 Reading Set

Blogger and post specific characteristics are likely to influence the amount of reading on a topic. Prior research suggests that individuals seek information from people with higher status who may have access to better source of information (Greve 2005). To account for the blogger status we construct a variable *HighPost* (*MiddlePost*) which indicates the number of posts written by *High* (*Middle*) status employees in the current period. The individuals were classified as *High*, *Middle* or *Low* based on their rank in the organization. For example, a representative *High* status individual in our study would be at the rank of a Chief- or a VP, whereas a *Middle* status would be at the rank of a senior manager, and the *Low* status would be at the rank of a developer.

The textual characteristics of the posts on a topic may affect the amount of reading on that topic. Two types of characteristics may be highly relevant here: *quality* and *sentiment*. We use text mining techniques to calculate measures for post quality in terms of *readability*, *comprehensiveness*, *grammatical errors*, and *sentiment* (*negativity* and *controversy*). *Readability* is measured as the Flesch Reading Ease Score which has a range of 0-100 with 0 meaning very hard and 100 very easy to read (Kincaid et al 1975). *Comprehensiveness* is measured as the length of post (Lu et al 2010). Sentiments expressed in the posts are classified as positive or negative using a Bayesian classifier (Clement and Sharp 2003) that is trained over word n-grams of order 5. The order was selected using cross-validation over a sentiment classification dataset (Pang and Lee 2004). Our classifier was initially trained over this dataset and then recalibrated using relevance feedback from the blog dataset. A *controversy* of zero indicates that reader opinion about the post content is undivided and its value away from zero indicates higher division or reader opinion. The grammatical error of the blog post was measured using the Link Grammar parser on each sentence of the post (Grinberg et al. 1995). The post textual characteristics are calculated for each post and then averaged across all posts on the topic in a given month. For more details on each of these measures please see the Appendix to this paper. Note that blogger and post specific variables in the reading set are constructed at a Topic-Month level.

We also account for posts that readers find interesting. Interests in the posts are measured by two variables: *PopPost*, and *ComPost*. *PopPost* measures the average popularity of the posts in terms of number of readings. *ComPost* is the number of comments made by readers to the post.

Table 2A: Definition of Variables

	Variable	Definition
Reading Set Corresponds to matrix Z in Sections 2.1, 2.2, and 2.3	$HighPost_{ct}$ ($MiddlePost_{ct}$)	Number of posts written in Topic c in month t by high (middle) status employees.
	$PopPost_{ct}$	Average popularity of posts in topic c in month t in terms of number of readings.
	$ComPost_{ct}$	Average Number of comments made by readers to posts in topic c in month t .
	$Readability_{ct}$	Average Flesch Reading Ease Score for posts in topic c in month t (0 very hard; 100 very easy).
	$Comprehensiveness_{ct}$	Average length of posts in topic c in month t .
	$Negativity_{ct}$	Average sentiment of posts in topic c in month t (1 all posts are negative; 0 all posts are positive).
	$Controversialness_{ct}$	Average standard deviation of the sentiment of the comments to posts in topic c in time t .
	$GramError_{ct}$	Average number of grammatical errors in posts in topic c in month t .
State Transition Set Corresponds to matrix X in Section 2.3	$ComPost_{it}$	Average Number of comments made by readers to posts read by reader i in month t .
	$Readability_{it}$	Average Flesch Reading Ease Score for posts read by reader i in month t (0 very hard; 100 very easy).
	$Comprehensiveness_{it}$	Average length of posts read by reader i in month t .
	$Negativity_{it}$	Average sentiment of posts read by reader i in month t (varies from 0-1; 0 all posts are positive; 1 all posts are negative).
	$Controversialness_{it}$	Average standard deviation of the sentiment of the comments to posts read by reader i in time t .
	$GramError_{it}$	Average number of grammatical errors in posts read by reader i in month t .
Reader Characteristics Set Corresponds to matrix D in Sections 2.2, and 2.3	$Gender_i$	Indicator variables that equals 1 if reader i is a male and 0 otherwise.
	$Asia_i$	Indicator variables that equals 1 if reader i is from Asia and 0 otherwise.
	$NAmerica_i$	Indicator variable that equals 1 if reader i is from North America and 0 otherwise.
	$Tech_i$	Indicator variable that equals 1 if reader i 's job includes technical work and 0 otherwise.
	$HighSi$	Indicator variable that equals 1 if reader i is high status employee and 0 otherwise.
	$MiddleSi$	Indicator variable that equals 1 if reader i is middle status employee and 0 otherwise.

3.1.2 State Transition Set

State transitions may be affected by situational factors (McAllister 1982) or an inherent desire for variety. While the individual specific inherent desire for variety would be captured through the individual specific random intercept, we can construct several situational variables. In the context of blog reading they would correspond to the reading experience of a reader. Hence, characteristics of the posts read by the reader in the present period can shape her experience and hence influence her switching. The post characteristics related variables are similar to the reading set variables (*readability*, *comprehensiveness*, *grammatical correctness*, and *sentiment* (*positive/negative* and *controversy*)). However, in this case they are constructed from posts that employees read in the prior period. The characteristics corresponding to the *quality* of the post would determine a reader's experienced quality. The effect of characteristics corresponding to the *sentiment* is not that obvious. *Negativity* and *controversialness* add variety to otherwise typically positive posts. At the same time, over exposure to them could be emotionally draining. Note that the variables in the State transitions set are calculated at Reader-Month level. Posts characteristics variables considered for state transition from time t to $t + 1$ are calculated from the reading behavior of the individual in period t .

3.1.3 Reader Characteristics Set

To control for reader specific effects we include several reader characteristics. We control for *status* specific effects by including two indicator variables representing High status (*HighS*) and Middle status (*MiddleS*) reader. Further, to account for differences that may exist among employees with technical expertise and others, we use a binary variable (*Tech*) which equals one if an employee is from a technical department and zero otherwise. We include an indicator variable *Asia* (*NAmerica*) which equals one when the reader is from Asia (*NAmerica*) and zero otherwise. Further, to account for gender specific effects, we construct an indicator variable, *Gender*, which equals one for male and zero otherwise.

Tables 2B and 2C provide descriptive statistics of Reading set and State transition set variables. These statistics indicate that on average most of the posts are written and read on NWE and NWK. WMK

posts receive on average more comments than posts on other topics. On average, WTK posts are harder to read than others, and NWE posts are longer than others. On average very small number of posts display negative sentiment across topics, and posts are moderately controversial across topics. WTK posts on average have higher number of grammatical errors than others. Table 2 C highlights some interesting statistics. On comparison of average negativity and controversialness of posts shown in Table 2B with the corresponding values in Table 2C, one can notice that readers are more attracted to negative and controversial posts. The values of the other characteristics are on average similar in Table 2B and 2C.

Table 2B: Descriptive Statistics of Reading Set Variables

		NWE	NWK	WMK	WTK
		Mean (Variance)	Mean (Variance)	Mean (Variance)	Mean (Variance)
<i>Reading Set</i> Corresponds to matrix Z in Sections 2.1, 2.2, and 2.3	<i>HighPost_{ct}</i>	86.42 (21.14)	141.71 (82.88)	95.24 (53.19)	53.26 (32.47)
	<i>MiddlePost_{ct}</i>	392.17 (114.19)	221.86 (97.94)	125.77 (44.51)	146.32 (119.42)
	<i>PopPost_{ct}</i>	44.61 (170.91)	31.18 (116.73)	12.48 (87.40)	10.72 (52.97)
	<i>ComPost_{ct}</i>	1.74 (4.11)	1.93 (3.82)	3.82 (9.54)	1.23 (8.41)
	<i>Readability_{ct}</i>	54.17 (14.26)	50.09 (19.93)	58.81 (11.19)	43.32 (21.68)
	<i>Comprehensiveness_{ct}</i>	50.91 (17.13)	28.53 (27.23)	24.85 (22.14)	39.52 (31.14)
	<i>Negativity_{ct}</i>	0.07 (0.13)	0.13 (0.08)	0.10 (0.11)	0.03 (0.06)
	<i>Controversialness_{ct}</i>	0.26 (0.18)	0.31 (0.09)	0.34 (0.13)	0.21 (0.08)
	<i>GramError_{ct}</i>	1.71 (3.01)	1.49 (2.73)	1.14 (2.18)	3.51 (2.85)

Table 2B: Descriptive Statistics of State Transition Set Variables

		Variables	Mean (Variance)
<i>State Transition Set</i> Corresponds to matrix X in		<i>ComPost_{it}</i>	2.07 (7.53)
		<i>Readability_{it}</i>	53.01 (24.19)
		<i>Comprehensiveness_{it}</i>	39.79 (23.92)
		<i>Negativity_{it}</i>	0.24 (0.10)
		<i>Controversialness_{it}</i>	0.38 (0.16)
		<i>GramError_{it}</i>	1.89 (2.87)

4 Model Estimation and Results

We performed several preliminary checks on the data. Initial investigations revealed that the post textual characteristics and their squares were highly correlated which raised concerns of multicollinearity. To reduce the correlations we mean centered the post textual characteristics which reduced the correlations to an acceptable level. The VIF for the transformed data is less than 2 which ensures that multicollinearity is no more a problem. Several of the reading set variables appear skewed which can lead to biased or non interpretable estimates. As suggested by Gelman and Hill (2007) we log transformed these highly skewed variables.

All three models of reading behavior are estimated using a standard MCMC hierarchical Bayes estimation procedure, using a Gibbs Sampler and the Metropolis-Hastings algorithm coded in Matlab (Rossi et al 2005). To reduce the autocorrelation between draws of the Metropolis-Hastings algorithm and to improve the mixing of the MCMC procedure, we used an adaptive Metropolis adjusted Langevin algorithm (Atchade 2006). In the hierarchical Bayes procedure for the dynamic model, the first 30,000 observations were used as burn-in and the last 25,000 were used to calculate the conditional posterior distributions. Convergence is assessed by inspecting the time series of draws (Gelman and Rubin 1992). The full estimation procedure is provided in an Appendix to this paper. We impose an identifying restriction on the dynamic model (McDonald and Zucchini 1997). The restriction enforces that the reading diversity (measured by Shannon diversity index) increases from State I to $S-I$. We introduce state S as a dormant state by imposing very small reading probability across topics to capture months when an employee completely stops reading. Further, the initial state distribution (π) for the dynamic model is derived analytically as $\pi = \pi \bar{\Phi}$, where $\bar{\Phi}$ is a transition matrix, which is calculated at mean values of covariates (McDonald and Zucchini 1997).

The number of states for the latent state and dynamic models is determined by comparing the BIC and log marginal density for models with different number of states (up to 6 states). The log marginal density is calculated as the harmonic mean of individual log likelihood across iterations (Newton and

Raftery 1994). The latent state (dynamic) model with four (five) states outperforms all other latent state (dynamic) models on the BIC as well as log marginal density criterion. Further the dynamic model with five states outperforms the simple model and the latent state model with four states. The BIC and log marginal density for the simple model, the latent state model with four states, and the dynamic model with five states are shown in Table 3. As the dynamic model explains the data best we will only discuss results from this model. Results from the other models are available from the authors.

Table 3: Model Comparison

Models of Reading Behavior	Number of States	BIC	Log-Marginal Density
Simple Model	1	3814484.9	-1906862.5
Latent State Model	4	2692818.5	-1345389.4
Dynamic Model	5	1940424.4	-966152.7

4.1 High Level Characteristics of the Estimated States

Table 4 presents the mean amount of reading in each topic in each state. The four states differ in their amount of reading on each topic. The Shannon Diversity Index and average amount of posts read in each state are also provided in the table. While the diversity of reading across topics increases as one moves from 1 to 5 by design, the intensity of the states decreases along the same path. That is, the more focused the readers are in their reading, the more they read. States 1, 2, and 3 have a very focused reading behavior where the employee reads a lot on one or two topics and very little on others. Whereas State 1 is focused more on non work related topics (NWE and NWK), State 2 is focused on NWK and WMK, and State 3 is focused on WTK primarily. In comparison, State 4 is quite diverse with similar amount of reading on each topic. The Shannon Diversity index also confirms this finding as State 4 has the highest diversity index. We refer to States 1, 2, and 3 as focused reading states, and State 4 as the exploratory reading state. The fifth state (State 5) is a dormant state where individuals do not read any topic.

Table 4: Mean Amount of Reading in a Topic in Each State

	Mean Amount of Reading				Shannon Diversity	Intensity
	NEW	NWK	WMK	WTK		
State1	44.07	11.18	0.76	0.59	0.59	56.60
State2	1.12	27.34	14.17	1.06	0.81	43.69
State3	1.78	1.59	5.66	17.46	0.90	26.49
State4	3.97	3.89	3.21	2.83	1.27	13.90

We begin with the coefficients corresponding to the reading set variables and then discuss those of the state transition set variables. Table 5a (5b) presents the estimated parameters for States 1 and 2 (3 and 4). The effect of most of the variables is similar in direction across topics and states. However, the effect sizes are different leading to different preference for the amount of reading across topics and states as shown in Table 4.

4.2 Reading Set Effects

The reading set effects are summarized in the Tables 5A and 5B. Let us examine the effects of post characteristics, blogger characteristics, and reader characteristics one at a time.

4.2.1 Post Characteristics

It comes as no surprise that posts with high readership have more comments. What is not so obvious, however, is that posts that display negative sentiment or are controversial are well read across topics and states. These factors appeal to the sentiment of the readers and are more likely to be discussed with others making them aware of the existence of the post. Further, moderately negative posts attract more readers possibly because they are more likely to offer a critical viewpoint on the matter than a simple praise or more effusive posts. However, excessively negative posts are more likely to contain pure criticism than a balanced analysis. The results show that very high negative posts do not lead to more reading.

Surprisingly, readability, grammatical correctness, and comprehensiveness do not affect the amount of reading on any topics and states. These characteristics do not appeal to the emotions of the readers and hence are less likely to be discussed with others. Hence, an individual would get to know these characteristics only after reading the post.

4.2.2 Blogger Characteristics

Posts on (non technical topics) NWE, NWK, and WMK written by high or middle status employees are read more compared to the ones written by low status employees across all states. However, on technical topic (WTK) only the posts written by middle status employees are read more than others. This is potentially a sign of readers being aware of the expertise of the bloggers. High status employees are primarily in managerial position and, hence, would provide better perspective on work related managerial knowledge posts. By contrast, a significant fraction of middle status employees have technical positions and hence would provide richer knowledge in their technical posts.

4.2.3 Reader Characteristics

Employee status affects the amount of reading across topics. Other than WTK, High and Medium status employees read more posts in all topics compared to low status employees across all states. Middle status employees read more than low status employees across all topics and states. It is interesting that the availability of time would have suggested the reverse relationship with low status employees reading the most followed by middle and high status employees. It appears that the blogs have been well adopted by high and middle status employees and to a lesser extent by low status employees. Technical employees read more NWE, NWK, and WTK posts than others across all states. Potentially, technical employees are more internet savvy and prefer to spend more time on internet than others. Hence they have more opportunities to access these blogs.

Employee location also affects the amount of reading. North Americans and Asians read more NWE and NWK posts than others in most of the states. North Americans read less and Asians read more than others WTK posts across most of the states. Our results also indicate that gender does not have any significant impact on the amount of reading in any Topic or state. There is also significant unobserved heterogeneity in the reading behavior of individuals as indicated by the significant $\text{Var}(\theta_i)$ across topics and states.

Table 5A: Hierarchical Bayesian Parameter Estimates of Dynamic Model of Reading Behavior

Variables	State 1				State 2			
	NWE	NWK	WMK	WTK	NWE	NWK	WMK	WTK
Reader Characteristics^Y								
<i>Technical</i>	4.81***	0.87**	-0.52	1.08***	0.71*	1.12**	-0.37	1.29***
<i>High Status</i>	1.79*	0.59**	1.07***	-0.17*	0.31*	0.78**	1.16***	-0.32**
<i>Middle Status</i>	3.98***	0.89*	0.94**	0.26**	1.07**	1.14**	0.99**	0.56***
<i>Individual Indicator</i>	26.14***	5.94***	-8.17***	-1.97***	-6.39***	18.55***	4.18***	-1.17***
$\mu(\theta_i)$	30.93***	7.18***	-7.29***	-1.51***	-5.19***	21.58***	5.93***	-0.31***
$Var(\theta_i)$	6.24***	4.17***	2.49***	1.16***	3.17***	4.84***	3.27***	2.37***
Blogger Characteristics								
<i>High Post</i>	2.29*	0.89**	2.32***	0.22	1.27	0.87***	2.44***	-0.07
<i>Middle Post</i>	2.15**	0.38**	0.48*	0.37**	0.75*	0.44**	0.36*	0.29**
Post Characteristics								
<i>PopPost</i>	0.76*	0.67*	0.62*	0.19*	0.29	1.07***	0.59**	0.21*
<i>ComPost</i>	2.94***	2.39***	3.07***	2.27***	0.32*	2.44***	3.11***	2.11***
<i>Readability</i>	1.75	1.52	0.17	0.03	0.54	1.63	0.15	0.11
<i>Comprehensiveness</i>	0.21	0.34	0.19	0.08	0.17	0.31	0.25	0.06
<i>NSentiment</i>	-0.28***	0.11*	-0.41**	-0.05*	-0.24***	0.14*	-0.36**	-0.03*
<i>NSentiment²</i>	-1.91***	-0.65**	-2.92***	-0.49**	-1.73***	-0.73**	-2.67***	-0.38**
<i>Controversy</i>	0.88***	0.49***	0.96***	0.16***	0.79**	0.59***	0.91**	0.22***
<i>GramError</i>	-0.15	-0.45	-0.27	0.13	-0.21	-0.36	-0.11	0.14

*** The 99% confidence interval does not include zero.

** The 95% confidence interval does not include zero.

* The 90% confidence interval does not include zero.

NSentiment is mean centered and *NSentiment²* is the square of mean centered *NSentiment*.

High Post, *Middle Post*, and *PopPost* are log transformed.

^Y Reader gender, and country specific variables though included in estimation are not reported here to conserve space.

Readability (*Comprehensiveness* and *GramError*) is (are) scaled down by a factor of 100 (10).

$\theta_i = \boldsymbol{\vartheta}'\mathbf{D}_i + \varepsilon_{\theta_i}$, where $\mathbf{D}_i = \{\text{Gender}_i, \text{Asia}_i, \text{NAmerica}_i, \text{Tech}_i, \text{HighS}_i, \text{MiddleS}_i, \text{Individual Indicator}_i\}$

Table 5B: Hierarchical Bayesian Parameter Estimates of Dynamic Model of Reading Behavior

Variables	State 3				State 4			
	NWE	NWK	WMK	WTK	NWE	NWK	WMK	WTK
Reader Characteristics^Y								
<i>Technical</i>	0.66*	0.54**	-0.75	3.99***	0.73***	0.71**	-0.54	1.17***
<i>High Status</i>	0.83*	0.36*	1.14***	-0.17**	0.35*	0.47*	0.09***	-0.11*
<i>Middle Status</i>	1.19**	0.62*	0.91**	0.68***	1.21***	0.75*	0.69**	0.82*
<i>Individual Indicator</i>	-5.34***	-4.82***	-1.77***	10.46***	-2.95***	-2.14***	-3.92***	-0.27***
$\mu(\theta_i)$	-3.16***	-3.22***	-0.98***	13.89***	-1.08***	-1.29***	-3.66***	1.39***
$Var(\theta_i)$	2.89***	1.06***	0.86***	2.09***	1.04***	1.44***	1.46***	0.99***
Blogger Characteristics								
<i>High Post</i>	0.21*	0.76*	1.82***	0.13	0.19*	0.73**	1.93***	-0.07
<i>Middle Post</i>	0.56	0.34*	0.29*	0.45**	0.53*	0.36*	0.37**	0.19*
Post Characteristics								
<i>PopPost</i>	0.18	0.77**	0.52*	0.77*	0.17*	0.72*	0.51*	0.23*
<i>ComPost</i>	0.39*	1.67**	2.55***	3.15***	0.38**	1.91***	2.53***	1.87***
<i>Readability</i>	0.40	0.84	0.16	0.17	0.37	1.07	0.16	0.13
<i>Comprehensiveness</i>	0.13	0.34	0.22	0.19	0.19	0.09	0.21	0.07
<i>NSentiment</i>	-0.19*	0.14*	-0.34**	-0.06*	-0.22***	0.18*	-0.41**	-0.11*
<i>NSentiment^2</i>	-1.65***	-0.68**	-2.79***	-0.51**	-1.83***	-0.79**	-2.47***	-0.78**
<i>Controversy</i>	0.66**	0.73**	0.83*	1.19***	0.76**	0.96***	0.76***	0.23*
<i>GramError</i>	-0.23	-0.45	-0.21	0.22	-0.07	-0.45	-0.12	0.17

*** The 99% confidence interval does not include zero.

** The 95% confidence interval does not include zero.

* The 90% confidence interval does not include zero.

NSentiment is mean centered and *NSentiment^2* is the square of mean centered *NSentiment*.

High Post, *Middle Post*, and *PopPost* are log transformed.

^Y Reader gender, and country specific variables though included in estimation are not reported here to conserve space.

Readability (*Comprehensiveness* and *GramError*) is (are) scaled down by a factor of 100 (10).

$\theta_i = \boldsymbol{\vartheta}'\mathbf{D}_i + \varepsilon_{\theta_i}$, where $D_i = \{Gender_i, Asia_i, NAmerica_i, Tech_i, HighS_i, MiddleS_i, Individual Indicator_i\}$

4.3 State Transition Set Effects

The state transitions set effects are more comprehensible when presented in a state transitions matrix format. Hence, we present all the state transition set effects in the form of transition matrices and discuss how changes in the state transition set variables induce change in the state transition probabilities. Table 6a presents a transition matrix that shows the baseline transition probabilities across states. The probabilities are constructed for a Low status, non Technical employee, from location other than North America and Asia, who has read posts that were at the mean level of *readability*, *comprehensiveness*, *sentiment*, *controversy*, and *grammatical correctness*. This transition matrix is very revealing. The values across the diagonal represent the stickiness (how likely is the individual to continue in the same state) of the states. Table 6a shows that state 2 is the stickiest, followed by 3, 1, 4, and 5. State 1 has higher intensity than State 2, whereas state 3 and 4 have lower intensity than state 2. This reveals that moderate intensity of reading is sustained for a longer time. Theoretically the literature on satiation supports this finding that moderate level of consumption reduces the rate at which satiation occurs (Redden 2008).

Table 6A: Baseline Transition Probabilities

$t \backslash t+1$	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>
<i>1</i>	42.76%	3.88%	2.45%	22.54%	28.37%
<i>2</i>	2.98%	76.17%	1.40%	8.61%	10.83%
<i>3</i>	4.25%	5.30%	70.61%	8.73%	11.10%
<i>4</i>	23.34%	24.05%	19.11%	15.49%	18.00%
<i>5</i>	16.20%	15.25%	13.13%	41.05%	14.37%

The substantial and significant non-diagonal transition probabilities highlight a *variety-seeking* behavior of blog readers. A reader frequently switches from reading one set of topics to another dynamically. Over the timeline of the study a reader would switch several times from one state to another. See Figure 2 for some examples. More interestingly, the probability of a reader switching from one focused state to another focused state is less than 5%. A reader stays in one of the focused reading state for sometime after which he temporarily switches to the exploratory or the dormant state and then

switches back to a focused state. The exploration/dormant state acts as a redistribution state in a way that a reader switches from one focused state to the exploratory or dormant state and then to the same or another focused state with almost equal probabilities.

The variables in the state transition set consist of reader specific and post specific variables. We will first discuss the impact of post specific variables on state transitions.

4.3.1 Post Specific Variables

Table 6B presents the transition probabilities with the same setting as in Table 6A (baseline probability) but with readings in time t having one standard deviation higher grammatical errors than the mean. The transition probabilities indicate that States 1 and 2 are affected the most by higher grammatical errors, whereas States 3 and 4 are affected minimally. Note that in states 1 and 2 the readings are mainly focused in non technical areas (NWE, NWK, WMK) whereas in state 3 they are mainly focused on technical topics (WTK). Hence, the results indicate that readers of non technical posts are more put off by grammatical errors than readers of technical posts. Further, grammatical errors significantly increase the transition to exploratory or dormant state from States 1 and 2.

Table 6B: Transition Probabilities for One Standard Deviation above Mean Grammatical Errors

$t \backslash t+1$	1	2	3	4	5
1	31.82%	3.69%	2.85%	26.74%	34.80%
2	4.46%	55.58%	2.55%	16.51%	20.91%
3	4.38%	5.60%	69.68%	9.01%	11.27%
4	22.32%	22.47%	20.20%	16.05%	18.96%
5	16.20%	15.25%	13.13%	41.05%	14.37%

Reading more comprehensive posts (one standard deviation higher than mean) reduces the stickiness of states 1 and 2 but increases stickiness of states 3 and 4 as shown in Table 6C. Note that higher comprehensiveness entails greater length of posts thus increasing the intensity of reading. Whereas states 1 and 2 involve a reasonably high level of intensity, more comprehensive posts heightens intensity further and hence reduce their stickiness. In comparison, States 3 and 4 are low in intensity as such and

more comprehensive posts push their intensity to moderate level where they become stickier. Further, more comprehensive posts make states 1 and 2 so intense that readers satiate and exhibit a higher probability to transition to the dormant state.

Table 6C: Transition Probabilities for One Standard Deviation above Mean Comprehensiveness

$t \backslash t+1$	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>
<i>1</i>	35.23%	4.22%	2.69%	24.80%	33.06%
<i>2</i>	3.38%	71.56%	1.65%	10.39%	13.02%
<i>3</i>	2.62%	3.42%	72.13%	10.78%	11.05%
<i>4</i>	23.19%	24.10%	19.09%	15.65%	17.98%
<i>5</i>	16.20%	15.25%	13.13%	41.05%	14.37%

Readers are also sensitive to the readability of posts. Readability has a curvilinear relationship with state transition probabilities. Moderate level of readability of posts is best for increased stickiness of the state and stickiness decreases as the readability moves to very low or very high levels. Table 6D shows the transition matrix with same setting as Table 6A with one standard deviation above the mean post readability. Posts that are harder to read (one standard deviation above the mean readability) reduces the stickiness of the states containing non technical issues (1 and 2). Further, posts that are harder or too easy to read also increase the likelihood of readers moving to the exploratory or the dormant state.

Table 6D: Transition Probabilities for One Standard Deviation above Mean Readability

$t \backslash t+1$	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>
<i>1</i>	33.92%	4.26%	2.74%	26.10%	32.99%
<i>2</i>	3.24%	71.97%	1.61%	10.36%	12.82%
<i>3</i>	4.22%	5.20%	70.71%	8.78%	11.06%
<i>4</i>	23.13%	23.77%	19.46%	15.58%	18.06%
<i>5</i>	16.20%	15.25%	13.13%	41.05%	14.37%

One standard deviation above the mean controversy of the posts increases stickiness across the focused reading states as shown in Table 6E. However, the effects are more pronounced in states 1 and 2 that contain non technical posts. This also reduces the probability of readers switching to the dormant

state. Conversely, negative sentiment of a post displays a non linear relationship with stickiness. Extreme values (too high or too low) of negative sentiment reduce stickiness of a state. As shown in Table 6F one standard deviation higher negative sentiment above mean reduce stickiness across the reading states. However, as before the effect is more pronounced in states with focus on non technical readings. Highly negative posts are quite off putting to readers as they are more likely to switch to the dormant state after reading such posts. Overall, posts with some amount of negativity up to a level would add variety to reading beyond which it would become too emotionally draining for the reader to continue reading.

Table 6E: Transition Probabilities for One Standard Deviation above Mean Controversialness

t\ t+1	1	2	3	4	5
1	51.94%	3.72%	2.04%	18.86%	23.44%
2	2.46%	80.62%	1.16%	7.02%	8.71%
3	4.21%	5.28%	70.83%	8.68%	11.00%
4	23.79%	26.04%	18.24%	14.79%	17.14%
5	16.20%	15.25%	13.13%	41.05%	14.37%

Table 6F: Transition Probabilities for One Standard Deviation above Mean Negativity

t\ t+1	1	2	3	4	5
1	37.61%	3.87%	2.69%	24.57%	31.26%
2	3.45%	71.62%	1.67%	10.31%	12.96%
3	3.92%	4.86%	67.60%	10.38%	13.24%
4	24.49%	24.94%	19.32%	15.06%	18.19%
5	16.20%	15.25%	13.13%	41.05%	14.37%

4.3.2 Reader Specific Variables

Reader status is found to impact the transition probabilities. High status employees in state 1, 2, and 4 (3 and 5) find the state more (less) sticky compared to low status employees (Table 6A vs Table 6G). Middle status employees find the focused states more sticky compared to low status employees (Table 6A vs Table 6H). However, they find the exploratory or dormant state to be less sticky than lower status employees. Overall high status people are less likely to switch to the dormant state as compared to low and middle status employees. This finding combined with the results in Table 5A and 5B (high status

employees read more than low status employees but less than middle status employees) provides further support for the argument that a moderate intensity of reading is more sustainable. Further, high status employees in State 1 have higher probability of staying there compared to low status employees, but high status employees in other states are more likely to transition to State 2 compared to low status employees. However, middle status employees have a more varied behavior. When in State 1 they are more likely to stay there or switch to State 2 in comparison to low status employees. When in States 2 and 3 they are more likely to stick there. However, when in States 4 and 5 they gravitate more towards State 1.

Table 6G: Transition Probabilities for High Status Employees

t\ t+1	1	2	3	4	5
1	47.91%	4.01%	2.16%	19.09%	26.83%
2	1.87%	81.80%	0.67%	7.13%	8.54%
3	3.7%	8.57%	67.91%	9.66%	10.16%
4	20.91%	31.83%	15.49%	16.12%	15.64%
5	14.79%	21.85%	11.87%	38.63%	12.86%

Table 6H: Transition Probabilities for Middle Status Employees

t\ t+1	1	2	3	4	5
1	51.36%	6.29%	0.75%	19.86%	21.74%
2	3.06%	78.88%	1.40%	7.83%	8.83%
3	3.65%	4.73%	74.79%	8.37%	8.54%
4	26.21%	22.79%	20.01%	14.10%	16.88%
5	19.67%	14.72%	13.32%	38.44%	13.86%

Employees from technical departments find the states 1, 3, and 4 stickier than non technical department employees (Table 6A vs Table 6I). Further, their probability to move to the dormant state is less than the corresponding probabilities for non technical department employees. Technical employees are also more prone to switch to State 1 or 3 compared to other employees. It appears that technical employees inherently like more WTK or NWE type posts.

Table 6I: Transition Probabilities for Technical Employees

$t \backslash t+1$	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>
<i>1</i>	53.01%	2.39%	12.70%	15.36%	16.47%
<i>2</i>	7.59%	69.86%	4.94%	7.98%	9.65%
<i>3</i>	2.65%	2.39%	86.75%	3.76%	4.45%
<i>4</i>	25.53%	20.49%	24.78%	13.87%	15.33%
<i>5</i>	16.42%	12.91%	23.06%	35.45%	12.16%

While gender does not have any impact on state transitions, employee location significantly impact state transitions. States 1 and 3 are stickier for employees in Asia than others. In comparison, State 2 is stickier for employees in North America than others. In general, employees in North America are more likely to transition to State 2, but employees in Asia are more likely to transition to State 1 and 3. Further, both the employees in North America and in Asia have less probability to switch to the exploratory reading state or the dormant reading than others.

4.4 Model Performance

To evaluate the predictive power of our model, we used the estimated results to predict, in the hold out sample, the amount of reading in each topic in each month. We employ the simple model and the four state latent state model as baseline models to illustrate the performance of our dynamic model. For all the models we calculated the Root Mean Square Prediction Error (RMSPE) which is given in Table 7. As is clear from the results the out of sample predictive ability of the dynamic model is superior to both the simple model and the four state latent state model.

Table 7: Model Predictive Power Comparison

Models of Reading Behavior	Number of States	RMSPE
Simple Model	1	41.8%
Latent State Model	4	22.5%
Dynamic Model	5	12.9%

5 Discussions

5.1 Key Findings and their Relation to Existing Literature

The first key finding of this study is that employees exhibit a *variety-seeking* behavior in reading blogs. They frequently switch from reading on one set of topics to another. This *variety-seeking* behavior bears some resemblance to the brand loyalty and switching behavior in the consumption of frequently purchased non-durable physical goods (see e.g., McAlister and Pessemier 1982, Givon 1984, McAlister 1982). In the case of physical goods, brand loyalty leads to purchase of same product over time due to inertia. As an individual consumes the same item repeatedly due to inertia, it can cause boredom when curiosity leads to the consumption of something new. Research in this area has focused on identifying which products display more loyalty or variety seeking behavior as these behaviors lead to different managerial implications. In the case of blogs even though a reader may read on the same topic she essentially reads distinct posts leading to some variety within the topic. Due to this feature of blogs, it is not obvious a priori as to what extent *variety-seeking* would be present in blog reading behavior. A priori one would expect that blog readers should display a high amount of inertia as consumption that includes a basket of products is argued to be inertial (Galak et al. 2009). However, as our results reveal *variety-seeking* is widespread among blog readers. It appears that the blog readers are inherently *variety-seekers*. Our results extend the *variety seeking* behavior to the context of virtual goods and highlight the importance of considering both consumer inherent behavior and product nature to analyze inertial or *variety seeking* behavior. Further, our results also highlight the importance of an exploratory state (sampling phase) as an intermediate step between the switching of a reader from one focused state to another which has not been studied in the traditional literature on *variety-seeking*. This behavior has similarities with the information foraging behavior of humans (Fu and Pirolli 2007, Miller and Remington 2004). One of the important concepts in the information foraging theory is that of information scent (Pirolli 2007). Individual's searching for information assess whether their paths exhibit cues to the desired outcome. They continue their search as long as they sense they are getting closer to the desired outcome.

If they feel the cues are weak they give up. The readings on a topic in the exploratory state act as cues for what the reader may expect if she focuses on that topic. If the reader likes it she intensifies reading on the topic or she gives up. This explains why readers rarely switch directly from one focused state to another. Readers need to explore a few posts on a topic to learn about the topic and only then they delve deeper into it.

Figure 2¹ shows the states of a few randomly selected employees . This along with the transition probabilities shown in the earlier section illustrates the *variety-seeking* behavior of blog readers and highlights the importance of the exploratory state in the reading behavior. After readers are satiated by reading in one of the focused states they switch to the exploratory state to find more interesting information to read. When they find something new that is more interesting they switch to a new focused state. Other times readers do not like anything and switch to dormant behavior. Sometimes after being in the exploratory state they go back to their previous state.

Figure 2. Individuals' states over time.

¹ The probability that an individual i is in State k at time period t is given as (Hamilton 1994):

$$\frac{\sum_{c_{it}} \pi_{c_{it}} \times L(N_{it} | \alpha, \beta_1 c_{it}, \dots, \beta_4 c_{it}) \times \prod_{z=1}^{t-1} [\varphi_{r=c_{it-1}c=c_{it}} \times L(N_{it} | \alpha, \beta_1 c_{it}, \dots, \beta_4 c_{it})]}{\sum_{c_{it}} \pi_{c_{it}} \times L(N_{it} | \alpha, \beta_1 c_{it}, \dots, \beta_4 c_{it}) \times \prod_{z=1}^{t-1} [\varphi_{r=c_{it-1}c=c_{it}} \times L(N_{it} | \alpha, \beta_1 c_{it}, \dots, \beta_4 c_{it})]}.$$

The second major finding is that the readership on a topic and the switching across states are induced by post characteristics. Our modeling framework allowed us to separate out two distinct impacts of post characteristics. One effect concerns attracting the readers (i.e., amount of readership on a topic) while the other entails retaining them (i.e., making the states sticky). The characteristics of the posts that appeal to the *sentiment* of a reader, such as controversy and negative sentiment, affect both attraction and retention. While controversy has a linear relationship with attraction and retention, negative sentiment displays an inverse U-shaped relationship with them. However, the *quality* of the post in terms of grammatical correctness, comprehensiveness, and readability affect only retention. While retention increases linearly with grammatical correctness, it has an inverse U shaped relationship with readability. Further, comprehensiveness increases retention of states with low reading intensity and decreases retention of states with high reading intensity. To our knowledge no other study separates the impact of textual characteristics on reader attraction and retention.

Ghose and Ipeirotis (2010) study the relationship between textual characteristics (specifically readability, grammatical correctness, and comprehensiveness) of a review and its helpfulness. Lu et al (2010) investigate the relationship between the review textual characteristics (specifically, readability and comprehensiveness) and the likelihood that the reader would consider the review trustworthy. Both trustworthiness and helpfulness of a review would have an effect on the retention of a reader. Hence, our results on retention and post textual characteristics (specifically readability, grammatical correctness, and comprehensiveness) are qualitatively consistent with the results of Ghose and Ipeirotis (2010) and Lu et al (2010). Aggarwal et al (2009) found a positive correlation between readership of a post and the negativity of the sentiment expressed there in. However, they have also observed that the positive correlation vanishes as the posts become increasingly negative. We find that the effects of textual characteristics are less pronounced for posts on technical topics than on non-technical posts. A possible explanation for this effect is that individuals care more about quality of the technical content when they read a technical post and less about the textual quality.

The third key finding of this study is that an individual's tendency to switch away from a set of topics increases with the number of posts recently read above a threshold, below which it decreases with the number of posts. This finding sheds new light on how the amount of consumption affects the desire to switch away from the item. Literature on *habituation* and *satiation* provide theoretical support for this finding (McAllister 1982). When an individual reads a small number of posts it takes some time before the reader gets habituated and hence is likely to persist with the behavior (read on the same set of topics). However, as the amount of reading increases it first allows inertia to set in when the reader gets habituated to read the same topic repeatedly. However, once the amount of reading increases beyond a limit she gets satiated/bored and tends to switch away seeking variety.

5.2 Managerial Implications

From our analysis and results, it is obvious that there are opportunities for a firm to influence the blog reading behavior of its employees. There are two questions that the firm needs to consider. First, when to influence? As our analysis revealed, employees transition from one focused state to another via an exploratory or dormant state. The best time for the firm to influence an employee's reading behavior is when the employee is in an exploratory or dormant state. Using our framework, the firm can probabilistically identify the state of a reader at any given time (refer to Footnote 1). Second, how to influence? As our results showed, properties of the text have significant effects on state transition. If the employee encounters poor quality work related posts she is more likely to switch away from reading such posts. However, the firm can influence the reader by suggesting her work related posts with high quality as they may increase readership as well as stickiness. Further, an employee who reads a large number of work related posts should not be recommended same type of posts. It is likely that such recommendations would cause the reader to satiate faster. Besides the textual characteristics, the blogger characteristics can influence the readership of a post. To increase readership of WMK posts, the firm should encourage its high status employees to write more posts on this topic, whereas the middle status employees should be encouraged to write more on both WMK and WTK posts.

The identification of *variety-seeking* behavior of blog readers poses both opportunities and challenges for a firm that intends to influence employee blog reading behavior. It is possible to switch a reader to a preferred state from a firm's perspective. In fact, our results show that even without any intervention employees frequently switch to preferred States 2 and 3. However, the same behavior also poses challenges. While a reader is likely to switch to a preferred state, she is as likely to switch away from that state. Overall, preferred states 2 and 3 are stickier than the non preferred state 1 which is quite positive from the employer's perspective.

The results from this study are also useful to the bloggers. Our results reveal that textual characteristics of a post significantly influence its readership and reader retention. Bloggers can use this information in fine tuning their posts to increase the loyalty of their followers. In addition, a firm can provide tools or feedback to help bloggers improve grammatical correctness and readability of their posts.

5.3 Contributions

Besides the managerial implications, our study makes several contributions. Ours is the first study to investigate blog reading dynamics. We identify the *variety-seeking* behavior of blog readers. Further, we show that their behavior is also induced by post, reader, or blogger characteristics. The modeling framework and the results from our study contribute significantly to our understanding of how individuals consume user generated content and provide guidelines to content generators to produce more popular content. The results highlight the dynamic aspects of individual behavior in consuming online content. They shed new light on improving personalized recommendations which is quite challenging in the context of online content delivery systems (Liu et al 2009 and Murthi and Sarkar 2003). Further, Web 2.0 technologies present a new dimension to the knowledge management process. Our study illustrates the effectiveness of one Web 2.0 technology (Blogs) for knowledge sharing. Additionally, it provides strategies for improving the usability of this medium for knowledge sharing.

The data structure and context of this study required us to make several methodological changes to traditional dynamics state space models. A methodological contribution of this study is to develop a Quad-variate Negative Binomial dynamic model to capture observed and unobserved inter-temporal as well as cross sectional heterogeneities in blog reading behavior. We extend the binary choice and univariate Negative Binomial dynamic models developed by, for example, Rabiner (1989), Hamilton (1989) and Singh et al (2009) to the context of multivariate count models. The endogenization of state transitions, the incorporation of covariates to determine state transitions, and the inclusion of Hierarchical Bayes individual level unobserved heterogeneity in both the amount of reading and state transitions extend the traditional HMMs and regime switching models. Our modeling framework helps in highlighting the opportunities for influencing reading behavior of employees to align it with firm goals.

6 Conclusions

In this study, we investigated the blog reading behavior of employees in an enterprise. An important aspect of our study is the investigation of the dynamics in the reading behavior of employees. To facilitate this we developed a dynamic state space model of blog reading behavior. The model was calibrated on a longitudinal dataset consisting of enterprise blog data of 1000 employees of a Fortune 1000 IT services and consulting firm. The results reveal that employees display a *variety-seeking* behavior where they frequently switch from reading on one set of topics to another. This finding also highlights the advantage of our modeling framework over traditional static models. Such behavioral dynamics cannot be identified by a static model. Our results also indicate that transitions from one focused state to another are interspersed by an exploratory or a dormant state. Further, the transitions are found to be induced by post, and reader characteristics as well as an inherent desire for variety.

Overall, this study contributes to our understanding of the viability of employee blogs as an efficient knowledge sharing medium. Our study shows that while enterprise blogs are facilitating work relevant knowledge sharing, firms have an opportunity to influence the reading behavior of employees to

realize the full potential of blogs for knowledge sharing. The model and results from this study can be used to devise strategies to influence employee blog reading behavior and to align their readings to firm goals. The blog reading dynamics estimation of this study contributes to the development of theoretically grounded understanding of reading behaviors of individuals in online settings and more specifically in communities formed around user generated content.

For estimation purposes we consider only four different topics following the classification of blogs by the firm. To ensure that the results are not driven by the number of topics considered, we estimated the model with more than four different topics (reclassified the subcategories to a larger set of topics). As one can expect the new model leads to a larger number of states. However, the model with more topics and states still displays a *variety-seeking* behavior and the effects of post, blogger, and reader characteristics on attraction and retention are qualitatively similar to those reported in the paper.

This study has its limitations. First, the model considers the blog posting process as exogenous. It is possible that bloggers may learn overtime and update their frequency of posting as well as post characteristics based on the changes in their readership. While an incorporation of blog creation process is out of the scope of this paper, future research should consider developing a model that incorporates this possibility. Second, while our dataset is sufficiently large to have findings generalizable to the context of the host firm, studies using datasets from other firms are needed to ensure that the findings are generalizable. Third, we have data on blog readings within the firm. However, employees may read external blogs too. We do not have access to such readings. While the incorporation of individual level unobserved heterogeneity should account for such effects, future research should consider collecting data on both internal and external blog readings for a more robust analysis. Finally, our reading is based on the information that a reader has accessed the blog. It may be the case that the reader did not read the blog at all or read it partially which we do not observe. Future studies should consider collecting such

information through surveys and questionnaires. Overall, these limitations open up exciting avenues for future research.

References

- Aaker, J., S. Fournier, A. Brasel. 2004. When Good Brands Do Bad. *Journal of Consumer Research*, 31, 1-18.
- Aggarwal, R., R. Gopal, R. Sankaranarayanan, P.V. Singh. 2009. Blog, Blogger, and the Firm: Can negative Posts by Employees Lead to Positive Outcomes?. *SSRN: 1532277*.
- Aggarwal, R., R. Gopal, A. Gupta, H. Singh. 2010. Putting Money where Mouth is: The Relation between New Venture Financing and Electronic Word-of-Mouth. *University of Utah Working Paper*.
- Adamic, L., N. Glance. 2005. The Political Blogosphere and the 2004 US elections: Divided they Blog. *International Conference on Knowledge and Data Mining*.
- Adar, E., L. Adamic. 2005. Tracking Information Epidemics in Blogspace. *Web Intelligence*.
- Allenby, G, J. Ginter. 1995. Using Extremes to Design Products and Segment Markets. *Journal of Marketing Research*, 32, 392-403.
- Arbous, A., J. Kerrich. 1951. Accident Statistics and the Concept of Accident-Proneness. *Biometrics*.
- Atchade, Y. 2006. An Adaptive Version for the Metropolis Adjusted Langevin Algorithm with a Truncated Drift. *Methodology and Computing in Applied Probability*. 8(2) 235-254.
- Blood, R., J. Rodzvilla. 2002. *We've Got Blog: How Weblogs are Changing our Culture*. Perseus Publishing, Cambridge, MA.
- Bolton, R., K. Lemon. 1999. A Dynamic Model of Customers' Usage of Services: Usage as an Antecedent and Consequence of Satisfaction. *Journal of Marketing Research*, 36, 171-186.
- Chevalier, J., D. Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43(3) 345-354.
- Clement, R., D. Sharp. 2003. Ngram and Bayesian classification of documents for topic and authorship. *Literary and Linguistic Computing* 18(4): 423.

- Efimova, L. 2007. Getting value from employee weblogs: A knowledge management approach. In H. Jezzard (Ed.), *Applying Web 2.0: Innovation, Impact and Implementation*, proceedings of Online Information 2007, 43. London.
- Farrell, H., D. Drezner. 2008. The Power and Politics of Blogs. *Public Choice*, 134(1) 15-30.
- Forman, C., A. Ghose, B. Wiesenfel. 2009. Examining the Relationship between Reviews and Sales: the Role of Reviewer Identity Disclosure in Electronic Markets. *Information System Research*. 19(3) 291-313.
- Fournier, S. 1998. "Consumers and Their Brands: Developing Relationship Theory in Consumer Research. *Journal of Consumer Research*, 343-373.
- Fu, W.-T., P. Pirolli. 2007. SNIF-ACT: A Cognitive Model of User Navigation on the World Wide Web. *Human Computer Interactions* pp 335-412.
- Furukawa, T., Y. Matsuo, I. Ohmukai, K. Uchiyama, M. Ishizuka. 2007. Social Networks and Reading Behavior in Blogosphere. *Int'l Conf on Weblogs and Social Media*.
- Galak, J., J. Kruger, G. Loewenstein. 2009. *Is Variety the Spice of Life? It All Depends on the Rate of Consumption*. CMU working paper.
- Gelman, A., J. Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.
- Gelman, A., D. Rubin. 1992. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*. 7(4) 457-472.
- Ghose, A., P. Ipeirotis. 2010. Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. Forthcoming in *IEEE Transactions on Knowledge and Data Engineering*.
- Givon, M. 1984. Variety Seeking Through Brand Switching. *Marketing Science*, 3(1) 1-22.
- Greene, W.H., D.A. Hensher. 2003. A Latent Class Model of Discrete Choice Analysis: Contrasts with Mixed Logit. *Transportation Research B*.
- Greve, H. 2005. Interorganizational learning and heterogeneous social structure. *Organizational Studies* 26 1025-1049.
- Grinberg, D., J. Lafferty, D. Sleator. 1995. A robust parsing algorithm for link grammars. *Proceedings of the Fourth Workshop on Parsing Technologies*.

Guadagni, P., J. Little. 1983. A Logit Model of Brand Choice Calibrated on Scanner Data. *Marketing Science*, 2 (3), 203-238.

Hamilton, J.D. 1989. A New Approach to the Economic Analysis of Non-stationary Time Series and the Business Cycle. *Econometrica* 57(2) 357-384.

Heckman, J. 1981. *Heterogeneity and State Dependence*, In S. Rosen ed. Studies in Labor Markets, Chicago: University of Chicago Press, 91-139.

Huang, Y., P.V. Singh, A. Ghose. 2010. *Show me the incentives: A dynamic Structural Model of Employee Blogging Behavior*. CMU Working Paper.

Huh, J., L. Jones, T. Erickson, W.A. Kellogg, R. Bellamy, J. Thomas. 2007. BlogCentral: the role of internal blogs at work. *CHI '07 extended abstracts on Human factors in computing systems*, 2007, 2447 – 2452.

Kincaid, J., R. Fishhourne, R. Rogers, B. Chissom. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. *Navy Research Branch Report*, 1975:8.

Leskovec, J., M. McGlohon, C. Faloutsos, N. Glance, M. Hurst. 2007. Cascading Behavior in Large Blog Graphs. *SIAM International Conference on Data Mining*.

Li, S., J. Liechty, A. Montgomery. 2002. Modeling Category Viewership of Web Users with Multivariate Count Models. Working Paper, Carnegie Mellon University, GSIA.

Lin, Y. -R., H. Sundaram, Y. Chi, J. Tatemura, B. Tseng. 2006. Discovery of blog communities based on mutual awareness. In *Third annual workshop on the weblogging ecosystem*.

Liu, D., S. Sarkar, C. Sriskandarajah. 2009. An Optimal Resource Allocation Policy for Personalization in Content Delivery Sites. *Information Systems Research*, Articles in advance.

Lu, Y., K. Jerath, P.V. Singh. 2010. The Emergence of Opinion Leaders in Online Review Communities. *CMU working paper*.

Lucier, C., J. Torsilieri. 1997. Why Knowledge Programs Fail. *Strategy and Business*, 4th quarter, 14-28.

McAlister, L. 1982. A Dynamic Attribute Satiation Model of Variety-seeking Behavior. *Journal of Consumer Research*, 9, 141-150.

McAlister, L., E. Pessemier. 1982. Variety Seeking Behavior: An Interdisciplinary Review. *Journal of Consumer Research*, 9, 311-322.

- McDonald, I., W. Zucchini. 1997. *Hidden Markov and Other Models of Discrete Valued Time Series*. Chapman and Hall.
- Miller, C. S., R. W. Remington. 2004. Modeling Information Navigation: Implications for Information Architecture. *Human Computer Interaction*, 19, pp 225–271
- Miura, A., K. Yamashita. 2007. Psychological and Social Influences on Blog Writing: An Online Survey of Blog Authors in Japan. *Journal of Computer-Mediated Communication*, 12(4).
- Mowbray, M. 2004. Setting Up an Online Community as an Enterprise Information Tool. *HP Laboratories Bristol*. Accessed: <http://www.hpl.hp.com/techreports/2004/HPL-2004-144.pdf>
- Murthi, B., S. Sarkar. 2003. The Role of the Management Sciences in Research on Personalization. *Management Science*, 49(10) 1344-1362.
- Nardi, B. A., D. Schiano, M. Gumbrecht, L. Swartz. 2004. Why We Blog. *Communications of the ACM*, 47(12) 41-46.
- Nakajima, S., J. Tatemure, Y. Hino, Y. Hara, K. Tanaka. 2005. Discovering Important Bloggers based on Analyzing Blog Threads. WWW, Shiba, Japan.
- Nelson, L., T. Meyvis, J. Galak. 2009. Enhancing the Television Viewing Experience through Commercial Interruptions. *Journal of Consumer Research*, 36 (2).
- Netzer, O., J. Lattin, V. Srinivasan. 2008. A Hidden Markov Model of Customer Relationship Dynamics. *Marketing Science*. 27(2) 185-204.
- Newton, M., A. Raftery. 1994. Approximate Bayesian Inference by the Weighted Likelihood Bootstrap. *Journal of the Royal Statistical Society*, B, 56 (1), 3-48.
- Oliver, R. 1997. *Satisfaction: A Behavioral Perspective on the Consumer*, New York: McGraw-Hill.
- Oliver, R. 1999. Whence Customer Loyalty. *Journal of Marketing*, 63, 33-44.
- Pang, B., L. Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization based on Minimum Cuts. *Proceedings of ACL*, 271-278.
- Pirolli, P. 2007. *Information Foraging Theory: Adaptive Interaction with Information*. Oxford University Press.
- Rabiner, L.R. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of IEEE* 77(2) 257–285.

Redden, J. 2008. Reducing Satiation: The Role of Categorization Level. *Journal of Consumer Research*, 34 624–34.

Rossi, P., G. Allenby. 2003. Bayesian Statistics and Marketing. *Marketing Science*. 22(3) 304-328.

Singh, P.V., N. Youn, Y. Tan. 2010. A Hidden Markov Model of Developer Learning Dynamics in Open Source Software Projects. Forthcoming in *Information Systems Research*.

Srinivasan V., R. Kesavan. 1976. An Alternative Interpretation of the Linear Learning Model of Brand Choice. *Journal of Consumer Research*, 3, 76-83.

Storey, J., E. Barnett. 2000. Knowledge management Initiatives: Learning from Failure, *Journal of Knowledge Management*, (4)2 pp. 145-156.