Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Using Social Network Activity Data to Identify and Target Job Seekers

Peter Ebbes, Oded Netzer

To cite this article:
Peter Ebbes, Oded Netzer (2022) Using Social Network Activity Data to Identify and Target Job Seekers. Management Science 68(4):3026-3046. https://doi.org/10.1287/mnsc.2021.3995

http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 68, No. 4, April 2022, pp. 3026–3046
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Using Social Network Activity Data to Identify and Target Job Seekers

Peter Ebbes,[a] Oded Netzer[b]

[a] Department of Marketing, HEC Paris, Jouy-en-Josas 78351, France; [b] Columbia Business School, Columbia University, New York, New York 10027

Contact: ebbes@hec.fr, https://orcid.org/0000-0002-6561-5319 (PE); onetzer@gsb.columbia.edu , https://orcid.org/0000-0002-0099-8128 (ON)

Received: June 20, 2018
Revised: February 17, 2020; September 6, 2020
Accepted: October 14, 2020
Published Online in Articles in Advance: June 16, 2021

https://doi.org/10.1287/mnsc.2021.3995

Abstract. An important challenge for many firms is to identify the life transitions of its customers, such as job searching, expecting a child, or purchasing a home. Inferring such transitions, which are generally unobserved to the firm, can offer the firms opportunities to be more relevant to their customers. In this paper, we demonstrate how a social network platform can leverage its longitudinal user data to identify which of its users are likely to be job seekers. Identifying job seekers is at the heart of the business model of professional social network platforms. Our proposed approach builds on the hidden Markov model (HMM) framework to recover the latent state of job search from noisy signals obtained from social network activity data. Specifically, we use the latent states of the HMM to fuse cross-sectional survey responses to a job-seeking status question with longitudinal user activity data, resulting in a partially HMM. Thus, in some time periods, and for some users, we observe a direct measure of the true job-seeking status. We demonstrate that the proposed model can predict not only *which users* are likely to be job seeking at any point in time but also *what activities* on the platform are associated with job search and *how long* the users have been job seeking. Furthermore, we find that targeting job seekers based on our proposed approach can lead to a 29% increase in profits of a targeting campaign relative to the approach that was used by the social network platform.

History: Accepted by Juanjuan Zhang, marketing.
Funding: P. Ebbes acknowledges research support from Investissements d'Avenir (ANR-11-IDEX-0003/LabexEcodec/ANR-11-LABX-0047) and the HEC Foundation.
Supplemental Material: The data files and online appendix are available at https://doi.org/10.1287/mnsc.2021.3995.

Keywords: Hidden Markov model • data fusion • Bayesian estimation • targeting customers • customer analytics

1. Introduction

The increased availability of data at the customer level (Wedel and Kannan 2016) allows companies to effectively target customers based on their individual characteristics (Matz and Netzer 2017), their location (Fong et al. 2015), or their past behavior (Trusov et al. 2016). Of particular interest to companies are customers' transition to and from unobserved states of behavior that may be of financial importance to the firm, such as expecting a child (Hill 2012), buying a house, going to college, unemployment, or job search. It is often during these periods of life transition that the customer may be open to marketing offerings (Bronnenberg et al. 2012) or may have a need for a particular product or service. For example, customers who will soon be buying a new house may be interested in mortgage offerings and are therefore attractive targets for a bank offering mortgage products. For such marketing problems, the firm may use its longitudinal activity data about its customers, possibly complemented by cross-sectional limited observations regarding the true state of some customers (e.g., collected via surveys) to infer these behavioral states for all customers in the current and in future time periods.

The objective of this research is to explore how a firm can leverage longitudinal activity data to infer the customers' latent states of behavior that are at the heart of the firm's business operation. Specifically, we investigate how an online social network platform with a substantial professional networking component[1] may use data about the activity of its users on the platform to identify which of the users are job seeking at any point in time. This is a key challenge for the platform, because most job seekers do not publicly announce that they are seeking for a job (Garg and Telang 2018). We demonstrate that job-seeking behavior can be inferred through how job seekers use the social network platform. For instance, relative to users who are not job seeking, job seekers may exhibit different forms of engagement on the social network platform such as updating their profile, more often searching for companies, or trying to grow their social

network by sending invitations to connect to other users. Furthermore, users who start searching for a job, may exhibit increased activity on the platform compared with their own past activity. However, without knowing the job-seeking status of at least a subset of the users, we cannot know to what extent the observed activity on the platform relates to job search.

In order to infer a user's job-seeking status, which is both latent and transient in nature, we use the hidden Markov model (HMM) framework. We combine two sources of information: (a) a large set of platform activities observed over time, such as the number of visits to the social network platform, profile updates, job searches, or invitations to connect with other users; and (b) the responses to a job-seeking status survey of a subset of the users at a certain point in time. To combine these two sources of information, we propose two ways to extend the traditional HMM to a partially HMM (PHMM), in which the latent states correspond to different levels of job seeking and are partially observed through the survey responses. In our models, each state is characterized by a multivariate set of activities on the social network platform. The PHMM provides a natural way to fuse the cross-sectional survey data with the longitudinal activity data. Specifically, we fuse the true job-seeking status for a subset of users at the time they respond to the survey into the likelihood of a traditional HMM, making their latent states partially observable at that time. As such, the PHMM is calibrated incorporating, possibly noisy, information about job-seeking status for some users at some points in time, allowing to infer the job-seeking states of all users in all time periods.

We demonstrate that the proposed model can infer and predict not only which members are likely to be job seeking at any point in time but also how long the members have been job seeking. Because of the size of the user base of the social network platform, only a small subset of users can be surveyed at a given time period. Hence, we demonstrate the ability of the proposed model to predict job search status both for out-of-sample time periods and for out-of-sample users. We compare the predictive ability of the PHMM to the predictive ability of two commonly used machine learning approaches: random forest (RF) and Lasso regression. We find that the PHMM out predicts both alternative methods. Furthermore, the machine learning–based approaches, possibly because of their somewhat static nature, fail to capture the timing in which the user transitioned to a job-seeking state. Going beyond identification of the job-seeking state, we demonstrate that targeting job seekers based on our proposed approach can lead to a 29% increase in response rates and profits relative to the approach that was used at the time of the data collection. These analyses highlight the managerial implications of accurately predicting job-seeking behavior.

The contribution of our research is twofold. Our primary contribution is substantive. We demonstrate how companies can leverage customers' activity data (e.g., clickstream or panel data) to infer the customers' latent behavior (e.g., job-seeking status), where the latent behavior, as inferred from the actual observed behavior, is of significant financial importance to the company. We show how targeting users based on our approach can lead to a substantial financial benefit. Specifically, in the context of job seeking, we uncover activities on the social network platform that are linked with latent job-seeking behavior, such as increased activity and strategic use of the user's social network. Furthermore, in targeting customers with transient latent behavior, such as job seeking, the timing of identifying the latent state transition is important. We show that our approach detects transitions from one latent behavioral state to another. Our secondary contribution is a methodological one. First, we demonstrate how one can naturally use HMMs, with relatively simple modifications, to fuse one or more snapshots of survey data, taking into account possible uncertainty in the survey response, into the sequence of longitudinal activity data through the latent state component of the HMM's likelihood function. The fusion of snapshots of survey data are important given the substantive problem, as the observed activities are only indirect proxies of the latent behavior of interest. Second, we demonstrate how HMMs can be adapted to one-to-many mappings between the job-seeking states and the observed activity (e.g., one job-seeking state where job seekers use the platform and a second job-seeking state where job seekers do not use the platform to job search). Third, most HMM applications in marketing leverage the latent states as means to capture and predict the dynamics of the state-dependent observed activity (e.g., donations in Netzer et al. 2008, churn and usage in Ascarza and Hardie 2013). However, this paper, like several HMM applications outside of marketing (e.g., Hamilton 1989), is focusing on the inference and prediction of latent state membership (i.e., job-seeking status) itself.

This paper is organized as follows. In the next section, we briefly discuss the relevant literature. In Section 3, we discuss our data and results from model-free analyses that motivate our modeling choices. Section 4 describes the proposed modeling approach. Section 5 presents the empirical results of how the proposed PHMMs capture users' job search status and the duration of job search. Section 6 demonstrates the managerial use of the model to target job seekers and the implied increase in profitability due to targeting based on the proposed model. Finally, we present the conclusions and discuss the limitations of our study in Section 7.

2. Literature Review

Our work builds on several streams of research. From a substantive point of view, our work relates to the identification of latent states of behavior from observed activity data, more specifically, to the identification of job-seeking states. From a methodological point of view, our work relates to work on data fusion approaches and HMMs. We briefly discuss these streams of research next.

2.1. Identifying Job Seeking

The U.S. staffing and recruiting industry was estimated at $151.8 billion in 2019.[2] One of the most important challenges for recruiting and job search firms is identifying who is job searching and when. Using survey data, Garg and Telang (2018) provide strong empirical evidence that people are spending a substantial amount of time searching for jobs on professional social networking platforms. They report that job searchers leverage professional social network platforms in several ways. They can (1) search for jobs posted or research potential companies and recruiters; (2) connect with friends or colleagues who may be aware of jobs or who may serve as leads or referrals; (3) connect with recruiters; and (4) be contacted by recruiters or employers. Accordingly, increased activity on the platform during one's job-seeking process may include more page visits, more searches, in particular more job searches, and connecting more often with (well connected) others. Additionally, a job seeker may wish to update her profile on the platform to attract connections from others. At the same time, Garg and Telang (2018) find that many recruiters turn to social networking platforms. For instance, they report that 94% of recruiters turn to the professional social network site LinkedIn. Consequently, users of online social networking platforms may be targeted and contacted by recruiters regarding potential job opportunities.

Job seekers often use social network platforms to foster the power of the network to assist them with finding a job (Stopfer and Gosling 2013). Additionally, the strength of the tie between the job seeker and their connections may be an important factor in the job search process. For example, according to Granovetter (1973), weak ties are likely to offer new information about possible jobs. Garg and Telang (2018), on the other hand, find that, in the context of online professional social networks, stronger, as opposed to weaker, ties were more effective in generating job leads, interviews, and job offers. These studies suggest that job seekers leverage their social network and that job seekers may wish to enhance their social network structure when they are searching for a job. In the context of our study, for instance, this could suggest that a job seeker will try to connect to more people, in particular, people who are outside their current professional network (e.g., outside their current company).

These studies highlight the importance of social network platforms in the job search ecosystem and the possible approaches that job seekers take to search for a job on these platforms. However, these studies are primarily based on survey data regarding job-seeking practices and are therefore limited in scope. To the best of our knowledge, no previous study used secondary data from users' activity on a social network platform to identify how job seekers use the platform at different stages of their job-seeking journey. In this study, we show how noisy signals embedded in a user's activity data may be used to infer whether that user is seeking for a job.

2.2. Approaches to Identify Latent States

The importance of and opportunity in identifying customers' latent states of behavior has been long recognized in marketing and related fields. Research has explored the ability to identify and target customers based on their latent preferences (Rossi et al. 1996, Hauser et al. 2009), their commitment to or relationship with the firm (Netzer et al. 2008, Ascarza and Hardie 2013, Romero et al. 2013, Schwartz et al. 2014, Ascarza et al. 2018), their price sensitivity (Zhang et al. 2014), their stage in the purchase funnel (Montgomery et al. 2004), their learning strategies (Ansari et al. 2012), and their portfolio of products (Schweidel et al. 2011). A common theme for these papers is that they include a latent space model (often an HMM) that captures the underlying behavioral or preference states. HMMs are useful in applications where the unit of analysis can dynamically transition among a set of latent states, but the actual state is only indirectly observable through a set of noisy signals. This setting perfectly matches our scenario in which the platform users are transitioning over time among different states of job-seeking behavior, but the platform does not directly observe the job-seeking status of its users. Instead, the platform observes a host of users' activities, which may provide a noisy signal of the users' job-seeking statuses. For example, a user who updates their profile and uses the job searching tool is providing a strong signal of searching for a job.

There are several important distinctions between our work and previous HMM applications in marketing. First, most of the aforementioned papers infer the nature of the latent states from the state-dependent activity only, whereas in this paper, we infer the states by fusing survey responses into the HMM likelihood that identify the true state for a subset of the users at a certain point in time. Netzer et al. (2008) validated the latent states of alumni-university relationships by comparing post hoc the inferred alumni states with responses of alumni to a customer relationship survey. In this paper, however, we propose to directly fuse such survey responses (with or without error) into the HMM likelihood function. In that sense, our work is

more closely related to the limited work on PHMMs in marketing, in which some of the states are fully observed. Romero et al. (2013) developed a PHMM to capture customer lifetime value. In their model some of the states are always observed (e.g., customer churn) and others are always unobserved (e.g., customer activity states). Similarly, Ascarza and Hardie (2013) use *two clocks* for usage and churn, where the churn state is fully observable every fourth time period, but use activities are affected by the latent HMM states in every time period. Our PHMM specification and modeling approach are different from these aforementioned studies because, in our case, all states are unobserved, except that for some users in some time periods the specific state of the user becomes observable (or observable with noise) through the user survey response. From a modeling perspective, the aforementioned PHMMs restrict the state-dependent behavior and/or certain transitions in the HMM to a fixed value, whereas our approach modifies the PHMM likelihood function by changing the transition into the "observed" state in the time period in which it is observed. Variations of PHMMs have been proposed in other fields, for instance, to model partially labeled training data in machine learning applications of natural language processing (Scheffer et al. 2001), to understand precipitation and rainfall activity (Thompson et al. 2007), or to identify users through typist keystroke dynamics (Monaco and Tappert 2018).

Second, in most marketing applications of HMMs the objective is to predict a certain outcome measure (e.g., purchase or website visit), where the latent states are used to capture the dynamics that governs the data generation of the outcome measures. In this research, we are not interested in predicting future outcome measures (e.g., future activity on the platform) but are instead interested in inferring and predicting the latent state itself (e.g., the job-seeking state). This approach is more similar to the use of HMMs in applications outside marketing, such as image recognition (Yamato et al. 1992), speech recognition (Rabiner 1989), or DNA detection (Eddy 1998).

2.3. Data Fusion

To identify the job-seeking state, we fuse responses to a survey into the HMM, which identifies (possibly with noise) the respondents' job-seeking status at the time of the survey. We use the survey-fused HMM to infer the job-seeking status of a larger set of users in any given time period. In other words, we fuse the information observed in the survey both cross-sectionally (to other users) and longitudinally (over time).

Data fusion is generally concerned with combining data from different sources. Statistically speaking, data fusion may be seen as a missing data problem. The basic idea behind data fusion is to capture the joint distribution of a collection of observed variables from two (or more) databases, in which a subset of the variables is observed for all observations. Given the subset of common variables, the fusion is based on the conditional joint distribution for the remaining variables across all observations. The most basic data fusion approaches are "hot-deck" procedures that impute the missing observations with information of individuals that have complete information on all variables and are similar on the joint observed variables to those with the missing information (Ford 1983). Kamakura and Wedel (1997) propose a statistical approach to tackle the problem of data fusion using a finite mixture approach and a factor analytic approach (Kamakura and Wedel 2000). Gilula et al. (2006) use a Bayesian approach to estimate a joint distribution using a set of variables that are common across units with missing observations. Qian and Xie (2014) propose a nonparametric Bayesian approach for data fusion. Other data fusion approaches have been proposed for specific marketing problems, such as the fusion of choice-based conjoint data with individual-level sales data to improve the estimation of consumer preferences (Feit et al. 2010) or fusing individual-level data with aggregate data (Feit et al. 2013). Bradlow and Feit (2018) provide an excellent review of data fusion modeling in marketing.

Our approach for data fusion is similar in spirit to the approach taken by Kamakura and Wedel (1997, 2000). Similar to Kamakura and Wedel, we also use a latent variable approach to fuse observed behavior with unobserved states. Our goal is to fuse survey data on job-seeking status observed in one (or multiple) time period(s) to other time periods of the users for whom survey responses are observed (*time sampling*, Kamakura and Wedel, 2000) and to all time periods for other users for whom no survey responses are observed (*subsampling*, Kamakura and Wedel 2000). Similar to the latent factor or the latent class in Kamakura and Wedel (1997, 2000), our approach uses the HMM latent states to fuse the partially observed survey data with the longitudinal platform activity data. However, unlike the static nature of the latent variable in the Kamakura and Wedel studies, our latent variable is dynamic such that we go beyond cross-sectional fusion and fuse information both cross-sectionally and over time. We propose two relatively simple modifications to the traditional HMM to fuse the survey responses, where the first assumes that the survey responses perfectly reflect actual job-seeking behavior and the second allows for error in the survey responses.

In any data fusion problem, one needs to consider the nature of the missing observations (Bradlow and Feit 2018, Kamakura and Wedel 2000). In our case, a random sample of users received a job-seeking survey

in the fifth month of a 14-month data window. Thus, the missing job-seeking status can be considered missing at random (MAR) (Kamakura and Wedel 2000) for the remaining months of the data period for those users that responded to the survey for the purpose of imputing their job-seeking status, because these missing observations were caused by the researcher design (i.e., the timing of the survey). However, for imputing the job-seeking status for individuals who were surveyed but did not respond, or those who were not surveyed, there could be a selection bias due to response bias (Wachtel and Otter 2013). Thus, the MAR assumption may not hold for imputing the job-seeking status for individuals that never responded to a survey. Because we only observe activity data for users who received and responded to the survey and not for users who did not respond nor for users who did not receive a survey in the first place, we cannot fully assess the extent to which the survey responses violate the MAR assumption in this study. However, we will show below using responses to a second survey that was administered at the end of the data period that the decision to respond does not appear to correlate with the job-seeking status nor with the employment status of the respondents.

3. Data Description and Model-Free Evidence

3.1. Monthly User Activity Data

We have a unique data set from a large online social network platform that has millions of users. Our data set contains monthly platform activity during the period of April 2010–May 2011 for a sample of 2,814 users who responded to a job-seeking survey (described later). These users were members of the platform and had at least 12 months of activity during the data period.[3] The data contain more than 60 types of user activities on the platform, such as whether the user sent or received an invitation to connect, the number of monthly page views and the type of page views (e.g., members' or companies' profile pages), how many company searches were made, how many times the user updated any part of the profile page, and so on. To keep the modeling effort manageable we select and collapse these activities into nine main variables measured at the monthly level: (1) whether the user used the job search tool (no = 0/yes = 1), (2) whether the user updated any aspect of their profile page (no = 0/yes = 1),[4] (3) how many pages the user viewed on the platform, (4) how many searches the user made using the platform's search tool (e.g., search for another member, search for a company, etc.), (5) how many invitations to connect the user received, (6) how many invitations to connect the user sent, (7) how many new connections the user formed, (8) how many connections the user's new connections had (on average), and (9) a dummy variable for whether the user connected more with users outside (= 1) or inside their current company (= 0). Because of the long-tailed nature of the continuous variables (variables 3–8), we log-transform these variables as $f(x) = \log(1 + x)$. When we introduce the continuous variables into our model, we use a type 1 Tobit model to account for the mass of observations at zero.[5]

Because of the firm's data collection approach at the time of the data collection period, some types of activity are observable for the entire 14-month period, whereas other types of activity are observable only for the first five months of the data period. Specifically, we observe variables 1–4 for the entire 14 months and variables 5–9 only for the first 5 months. Such imbalance in data collection is quite common in firms' databases (Zarate et al. 2006). In the model section, we describe how we handle this data imbalance.

3.2. Job Search Survey Data

In addition to the monthly activity data, we also used the platform to survey the users in our sample at two periods in time regarding their job-seeking status. The first survey took place in month 5 of the data period (August 2010) and the second survey took place shortly after the last month of the data window (June 2011). We will fuse the first survey (hereafter the survey) into the model to identify the job-seeking states and hold out the second survey for validation (hereafter the validation survey). Clearly, it is impractical for the company to survey all of its users every month regarding their job-seeking status. Hence, an important part of this study is to develop an approach to fuse the survey responses with the social network platform activity data across users and over time to identify the latent job-seeking status of all users over time.

To maximize compliance, the job-seeking surveys were very short with only a few questions. The main question asked was "How would you classify your current job search status?" with the following response categories[6]:

1. I am completely happy in my current job and am **not interested in discussing** any new job opportunities,
2. I am not looking for a new job, but **would discuss** an opportunity with a recruiter to see if the job is meaningful,
3. I'm **thinking about** changing jobs and have reached out to close associates but am not actively looking,
4. I am **casually looking** for a new job two to three times per week or to test the market, and
5. I am **actively looking** for a new job and sharing my resume.

The second column in Table 1 shows the proportion of responses to each of the job-seeking categories in the survey. Approximately 21% (=11% + 10%) of the respondents are actively or casually looking for new

job opportunities, whereas 21% are not looking for new opportunities.[7]

3.3. Model-Free Evidence

3.3.1. Relationship Between Job Seeking Status and Activity During the Month of the Survey. In Table 1, we report the users' activity on the platform during the month of the survey by the users' responses to the job-seeking survey question. One of the activity variables we observe is whether the user used the platform's job search tool. A naïve approach to identify the latent state of job search would be to classify users that actually use the job search tool in a given month as active job seekers. The third column in Table 1 reports the proportion of users who use the job search tool during the month of the survey by their survey response category. As expected, we find that the job-seeking status survey response significantly correlates with the use of the job search tool ($\chi^2(4,\ N = 2814) = 227.97,\ p < 0.001$). Specifically, those who are actively looking for a job use the tool considerably more than other users. However, 52% of those who actively search for a job according to their survey response, and 75% of those who casually search for a job, did not use the job search tool during the month of the survey. Thus, although job seekers use the job search tool more frequently than those who do not search for a job, many job seekers cannot be identified with this single activity.

Examining other user activities, we find that in the month of the survey, active job seekers view, on average, more than twice as many pages on the platform as the other users ($F(4, 2809) = 26.98$, $p < 0.001$), search twice as often ($F(4, 2809) = 24.55$, $p < 0.001$), and have a higher probability to update their profile page ($\chi^2(4,\ N = 2814) = 65.50$, $p < 0.001$). We also observe that job seekers grow their social network differently from nonjob seekers. Users who indicate in the survey that they are job-seeking form more connections on the platform during the month of the survey than other users ($F(4, 2809) = 5.34$, $p < 0.001$). In addition, we find that job seekers were more likely to send invitations to connect, trying to expand their network ($F(4, 2809) = 10.42$; $p < 0.001$); however, they are not more attractive for other users to connect to, receiving no more or even fewer invitations to connect than other users ($F(4, 2809) = 1.04$, $p = 0.38$). Thus, there is an asymmetry between invitations sent and invitations received across the various job-seeking categories. Last, one could ask whether users strategically expand their network for job search purposes. To investigate this, we examine whether the five types of job seekers differ with respect to the type of users they try to connect to. We find that active job seekers seem to be strategic in growing their network, connecting to other users that have relatively more connections than the

Table 1. Comparison of the User Activity During the Month of the Survey Across Job Search Survey Responses

	Variables available for 14 months						Variables available for 5 months				
	Proportion		Average			Proportion	Average				
	Survey response	Uses job search tool (0/1)	Profile updates (0/1)	Page views	Total searches	More invitations outside than inside company	Invitations sent	Invitations received	Connections formed	Connections of invitee
1. Not interested	0.21	0.06	0.25	21.98	1.45	0.77	1.11	1.23	2.56	29.06
2. Would discuss	0.43	0.09	0.22	25.33	1.59	0.75	1.19	1.36	2.83	41.88
3. Thinking about	0.14	0.18	0.24	27.88	2.00	0.76	1.17	1.43	2.82	44.34
4. Casually looking	0.10	0.25	0.36	36.57	2.33	0.88	1.32	1.32	2.81	24.38
5. Actively looking	0.11	0.48	0.48	74.19	5.06	0.79	2.39	1.23	3.93	63.61
Test statistic (H0: no difference between groups)	1,009.90	227.97	65.50	26.98	24.55	3.38	10.42	1.04	5.34	2.82
p value	<0.001	<0.001	<0.001	<0.001	<0.001	0.50	<0.001	0.38	<0.001	0.02
N	2,814	2,814	2,814	2,814	2,814	398[a]	2,814	2,814	2,814	1,081[a]

Note. Absolute numbers for activity are scaled by an unknown number.
[a]The sample sizes for these variables are smaller because these variables are only observable when a user sent an invitation to connect. We only observe whether the user sent more invitations outside or inside its current company when for both users the current company field is observed.

users to whom the other job-seeking type users are connecting to ($F(4, 1076) = 2.82$, $p = 0.02$).

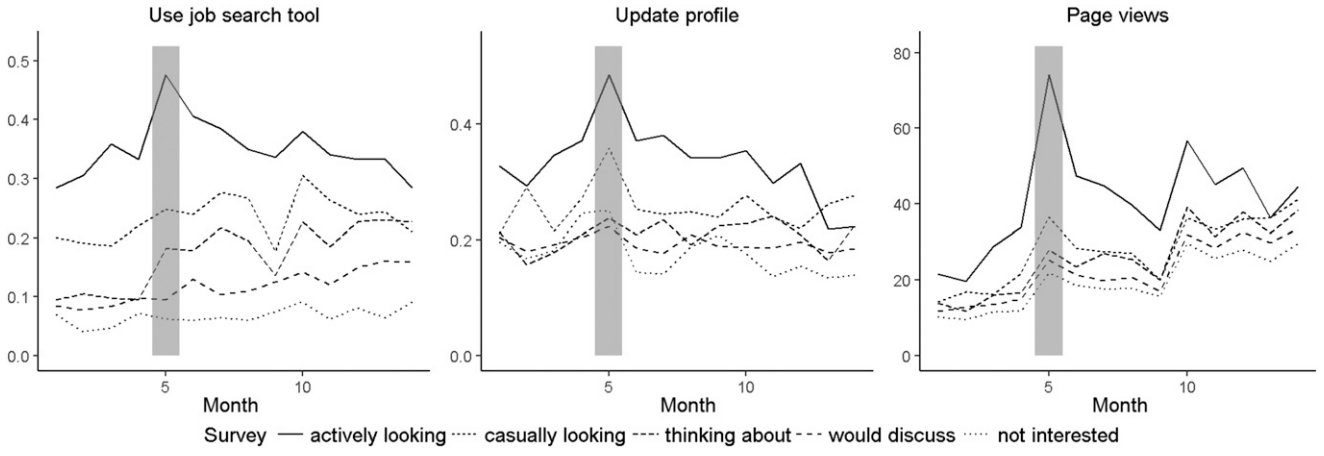
3.3.2. Longitudinal Analysis of Relationship Between Job Seeking Status and Activity. The analysis described previously provides a snapshot of the different user activities during the month of the survey. On the one hand, we find that job seekers exhibit different behaviors on the platform both in terms of platform activity, as well as in terms of social network activity. On the other hand, it seems that any one single activity cannot accurately reveal the user's job-seeking status. Hence, a multivariate approach to characterize job-seeking behavior is more appropriate. An additional source of information to infer job-seeking status may come from the users' longitudinal activity, as job seekers likely change their activity patterns over time, possibly even before starting their job search.

Figure 1 summarizes the time series of three of our main activity variables, along with the time stamp (shaded area) of the survey in the fifth month of the data period. The lines represent the level of average activity over time for the different users based on their response to the job-seeking survey question in month 5. That is, given the responses in month 5, we compute the average activity level in each month by the response categories of the job-seeking survey question. This allows us, for instance, to examine what those who reported to be active job seekers in the survey in month 5 did, on average, in the months before and after month 5. If longitudinal data are useful in predicting job seekers, we should expect an increase in average activity for users who state they are job seeking in the month of the survey but not for users who are not job seeking in the month of the survey. Furthermore, we may expect that most users who are active job seekers in month 5 find a job at some point, so their average activity likely decreases after month 5 and eventually returns to similar levels as for those who reported to be not seeking.

Several observations regarding Figure 1 are noteworthy. First, we observe that activity on the platform is increasing over time, reflecting the general growth of the social media platform. Particularly, the average number of page views and the use of the job search tool increase over time, reflecting the general growth of the social media platform. To account for such an increase, and to distinguish it from job search patterns, we include the number of unique visitors to the platform[8] during the data period as a covariate in our main model. Second, we find that changes in activity over time may be indicative of job-seeking status. For instance, the likelihood of updating the profile page peaks in month 5 for users who report to be active or casual job seekers but not for other users who report to be not job seeking in month 5. The increase in profile update activity seems to start before month 5, as some of these job seekers may have been searching for a while or may have been preparing their "window dressing" for the job search. As we move away from the survey month, the average activity level of those who report to be job seeking converges to the average activity level of the other users, as these users most likely have found a job by that time.

In sum, there are two important insights from the model-free evidence for building our model. First, job seekers exhibit different behaviors on the platform than nonseekers, and these behaviors should be characterized by a multivariate set of activities. A single observed activity cannot fully characterize the unobserved behavior (job search) of interest. Second, the activity levels of job seekers change over time, presumably when their (latent) job-seeking status changes. Thus, the users' activity levels and their

Figure 1. Average Monthly Activity Levels of Probability to Use the Job Search Tool, to Update the Profile, and the Number of Page Views During the Observation Period



Notes. The survey was fielded in month 5 (shaded area). Absolute numbers for activity are scaled by an unknown number.

change over time can be indicative of the users' latent states of job search. This setting is a natural case for a latent state model, such as an HMM, to identify job seeking from a set of multivariate activities. As the company cannot survey all users in all time periods, our proposed model needs to fuse the information from one or more surveys for a sample of users in one or more time periods. In the next section, we discuss our modeling approach.

4. Modeling Approach and Estimation

HMMs have been widely used to model latent states of behavior or latent states of the world (for a recent review of HMMs in marketing, see Netzer et al. 2017). This class of models suits our substantive research problem and data well, because we observe users' activities on the platform, which serve as noisy signals of the latent variable of interest: the users' job-seeking status. Furthermore, it is important to model the dynamics in the job-seeking states, because users transition in and out of different job-seeking statuses over time. We start with describing an approach that fuses the survey responses assuming they are a perfect representation of job search behavior and then discuss an approach that relaxes that assumption assuming imperfect survey responses.

4.1. A PHMM of Job Seeking with Data Fusion for the Survey Responses

We consider a HMM with K latent states. The latent state variable S_{it} takes on the values $\{1,2,\ldots,K\}$, capturing in which state user $i = 1,2,\ldots,N$ is in month $t = 1,2,\ldots,T$. We note that the number of states K need not directly match the number of job-seeking status survey categories because the same job-seeking status may correspond to multiple observed behaviors on the platform. For example, there could be two types of nonjob seekers on the platform, those who use the platform actively but not for job seeking and those who do not use the platform frequently other than occasionally logging in. Our proposed PHMM allows for this flexibility. We observe multivariate user activity data, $\boldsymbol{Y}_{it}$, where $\boldsymbol{Y}_{it}$ is a $P \times 1$ vector of P user activities (e.g., profile updates, total number of searches). In an HMM, we assume that the probability distribution of $\boldsymbol{Y}_{it}$ depends on S_{it}. For example, users in the active job-seeking state may be more likely to use the job search tool or view more pages relative to users who are not in the active job-seeking state.

Importantly, we observe the true job search status for some users in some time periods through their response to the job-seeking survey. Hence, the survey partially reveals the unobserved state S_{it} in the month of the survey, and we can use this information to update the likelihood function corresponding to the path $\{\ldots, S_{it-1}, S_{it}, S_{it+1}, \ldots\}$ taken. As we will show, the HMM framework provides a natural way to fuse the survey responses into the likelihood function. Fusing the survey responses into the HMM likelihood function helps to calibrate the latent states. At the same time, it facilitates anchoring the meaning of the latent states to the context of job search. The resulting modeling framework is a PHMM, rather than a traditional HMM framework, because the latent states are partially observed through the one-time survey response. We note that this representation of a PHMM is different from common PHMMs in marketing (Ascarza and Hardie 2013, Romero et al. 2013), because these models assume that some states are always observable for all users (e.g., a churn state), whereas other states are never observable. However, our model assumes that all states are partially observable for a subset of the users and only during certain time periods. The resulting formulation of the PHMM proposed here is therefore different from the one proposed in the above papers.

We build on a standard HMM commonly used in the literature (e.g., Netzer et al. 2008). The model consists of three main components: (a) the $K \times 1$ vector of initial state probabilities $\boldsymbol{\pi}_i = \{\pi_{i1}, \pi_{i2}, \ldots, \pi_{iK}\}$, (b) the $K \times K$ transition probabilities matrix $Q_i = \{q_{i11}, q_{i12}, \ldots, q_{i1K}, q_{i21}, q_{i22}, \ldots, q_{i2K}, \ldots, q_{iK1}, q_{iK2}, \ldots, q_{iKK}\}$, and (c) the $K \times K$ diagonal matrix M_{it}, which contains the state-dependent activity distributions m_{itj}, that is, $M_{it} = Diag\{m_{it1}, m_{it2}, \ldots, m_{itK}\}$. The users are likely to be heterogeneous in terms of their activity on the platform and in their approach to job search. We account for unobserved user-level heterogeneity by including random-effect intercepts in each of the three main components ($\boldsymbol{\pi}_i, Q_i$, and M_{it}). Including random-effect intercepts allows us to separate within-user baseline activity and transient job-seeking behavior. To the extent that different types of job seekers inherently exhibit different levels of activity on the platform, these should be captured by the user-specific intercepts.

4.1.1. State-Dependent Activity Distribution.

In our model, the state-dependent activity distribution is a multivariate distribution describing users' multiple activities on the platform. Conditional on the user's state S_{it}, we have a standard probability model for multivariate behavior. We model the discrete activities using a binary logit model. The continuous activities were log-transformed as $\log(1 + X)$ to capture the long tail observed in these activities and modeled as a type 1 Tobit regression model to capture the mass and bound at zero (Amemiya 1984). The probability distribution for the discrete variable p, $p = 1, 2, \ldots, P_1$, is defined as follows:

$$P(Y_{itp} = 1 | S_{it} = k, \boldsymbol{\theta}) = \frac{\exp(\delta_{0pk} + \delta_{1p} Z_t)}{1 + \exp(\delta_{0pk} + \delta_{1p} Z_t)}, \quad k = 1,2,\ldots,K, \tag{1}$$

where δ_{0pk} is the logit intercept for activity p given state k, and δ_{1p} is the regression coefficient for the

control variable Z_t (the unique number of visitors to the platform to capture general aggregate trends in activity during the data period). Similarly, the probability distribution for the continuous variable p, $p = 1, 2, \ldots, P_2$, is defined as follows:

$$f\left(Y_{itp}|S_{it}=k,\boldsymbol{\alpha}_i^M,\boldsymbol{\theta}\right)=Tobit_{type1}(\mu_{itpk},\sigma_{pk}^2), \tag{2}$$

with

$$\mu_{itpk}=\beta_{0pk}+\beta_{1p}Z_t+\alpha_{ip}^M, \tag{3}$$

where β_{0pk} is the intercept of the p th variable in state k, $p = 1, 2, \ldots, P_2, \beta_{1p}$ is the effect of the time trend on the p th variable, and α_{ip}^M is a user specific random intercept for the p th activity variable that captures the difference between user i's baseline activity and the population mean. The variance σ_{pk}^2 is the variance of the residual error term in the type 1 Tobit model for activity variable p and state k. In (2), $\boldsymbol{\alpha}_i^M$ represents the user-specific vector of random intercepts, and in (1) and (2), $\boldsymbol{\theta}$ represents a vector of fixed-effect parameters. Overall, the conditional probability of observing user i's multivariate platform activity at time t, given the user's latent state S_{it}, is given by the joint probability

$$\begin{aligned} m_{itj}&=P\left(\boldsymbol{Y}_{it}|S_{it}=j,\boldsymbol{\alpha}_i^M,\boldsymbol{\theta}\right)\\ &=\left(\prod_{p=1}^{P_1}P(Y_{itp}|S_{it}=j,\boldsymbol{\theta})^{Y_{itp}}\left(1-P\left(Y_{itp}|S_{it}=j,\boldsymbol{\theta}\right)\right)^{(1-Y_{itp})}\right)\\ &\times\left(\prod_{p=P_1+1}^{P_1+P_2}f(Y_{itp}|S_{it}=j,\boldsymbol{\alpha}_i^M,\boldsymbol{\theta})\right). \end{aligned} \tag{4}$$

4.1.2. Initial State Distribution and Transition Probability Matrix. We model the initial state distribution $\boldsymbol{\pi}_i = \{\pi_{i1}, \pi_{i2}, \ldots, \pi_{iK}\}$, as a multinomial logit model with K options, with parameters τ_j as baseline logit thresholds, α_{ij}^{π} as individual-level threshold random effects, $j = 1, 2 \ldots, K-1$, and $\pi_{iK} = 1 - \Sigma_j \pi_{ij}$. Similarly, we model each row $\boldsymbol{q}_{ik} = \{q_{ik1}, q_{ik2}, \ldots, q_{ikK}\}$ of the $K \times K$ transition matrix Q_i using a multinomial logit model, with ϕ_{kj} as the baseline intercepts for the logit probability that a user is transitioning from state k to state j in a given time period, and including an individual-specific random effect, a_{ikj}^Q, for $j = 1, 2, \ldots, K-1$, $k = 1, 2, \ldots, K$, as follows:

$$q_{ikj}=P\left(S_{it}=j|S_{it-1}=k,A_i^Q,\boldsymbol{\theta}\right)=\frac{\exp(\phi_{kj}+a_{ikj}^Q)}{\sum_{l\in K}\exp(\phi_{kl}+a_{ikl}^Q)}, \tag{5}$$

with $q_{ikK} = 1 - \Sigma_j q_{ikj}$.

4.1.3. Likelihood Contribution for User *i*. Ignoring for the moment that we observe job-seeking status survey responses, the probability of observed data for user i, given the user-specific vector of random intercepts $\boldsymbol{\alpha}_i$ and the vector of fixed-effect parameters $\boldsymbol{\theta}$, is given by

$$P(\boldsymbol{Y}_{i1},\boldsymbol{Y}_{i2},\ldots,\boldsymbol{Y}_{iT}|\boldsymbol{\alpha}_i,\boldsymbol{\theta})=\boldsymbol{\pi}_i M_{i1}Q_iM_{i2}Q_i\ldots Q_iM_{iT}\boldsymbol{\iota}, \tag{6}$$

where $\boldsymbol{\iota}$ is a $K \times 1$ vector of ones. The vector $\boldsymbol{\alpha}_i$ contains the user specific random intercepts for $\boldsymbol{\pi}_i, Q_i$, and M_i, that is, $\boldsymbol{\alpha}_i = (\boldsymbol{\alpha}_i^{\pi}, \boldsymbol{\alpha}_i^Q, \boldsymbol{\alpha}_i^M)$, where $\boldsymbol{\alpha}_i^{\pi} = (\alpha_{i1}^{\pi}, \alpha_{i2}^{\pi}, \ldots, \alpha_{iK-1}^{\pi})'$ is a $(K-1) \times 1$ vector, $\boldsymbol{\alpha}_i^Q = vec(\mathrm{A}_i^Q)$, A_i^Q is a $K \times (K-1)$ matrix with (k, j) th element α_{ikj}^Q, and $\boldsymbol{\alpha}_i^M$ is a vector of random intercepts for the continuous activity variables.[9] We assume a multivariate normal distribution for the upper-level model of the random intercepts, $\boldsymbol{\alpha}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}})$.

4.1.4. Fusion of Survey Responses with No Response Error (*Deterministic* Fusion PHMM). Next, we describe how to fuse the survey responses into the likelihood of the HMM to help identify the underlying latent states, resulting in a PHMM. Intuitively speaking, if user i responds to the job-seeking survey in time period t, then the paths of the latent state for time periods $t-1, t$, and $t+1$ are partially known. For example, if the user indicates she is in job-seeking state s in time period t, then only transitions into state s are allowed from time period $t-1$ to time period t. This will constrain the transition probability matrices for this user going into time period t. However, as mentioned earlier there could be multiple HMM states of user activity on the platform that correspond to the same job-seeking status; thus, the survey may not fully reveal which HMM state the user is in at the month of the survey. Additionally, the survey responses may include response error or response bias, when respondents fail to provide an accurate answer to the job-seeking question. We first assume no response error in the survey response (which we call *deterministic* fusion *PHMM*) and allow for possible survey response error (which we call *stochastic* fusion *PHMM*) in the next section.

To capture the partial observability of states during the months of the survey, we define $Q_{i,\cdot\to s}^t$ as a $K \times K$ matrix of zeros except for the s th column(s), which is the *sth* column(s) of Q_i. For example, consider a six-state HMM where the first state represents the non–job-seeking status, the second, third, and fourth states correspond, respectively, to the *would discuss*, *thinking about*, and *casually looking* job-seeking statuses, and the fifth and sixth states represent the active job-seeking status (e.g., one state may be characterized by low and the other state by high levels of platform activity). Suppose user i indicates she is an active job seeker in time period t. Now we allow user i in period t to only

transition into states $s = 5$ or $s = 6$ by constraining Q_i in period t as follows:

$$Q^t_{i,\cdot\to s=5,6} = \begin{bmatrix} 0 & 0 & 0 & 0 & q_{i15} & q_{i16} \\ 0 & 0 & 0 & 0 & q_{i25} & q_{i26} \\ 0 & 0 & 0 & 0 & q_{i35} & q_{i36} \\ 0 & 0 & 0 & 0 & q_{i45} & q_{i46} \\ 0 & 0 & 0 & 0 & q_{i55} & q_{i56} \\ 0 & 0 & 0 & 0 & q_{i65} & q_{i66} \end{bmatrix}. \quad (7)$$

If the observed job-seeking status is characterized by only one PHMM state, then only one column in the transition matrix in Equation (7) would be set to probabilities and the rest to zeros. The likelihood function in Equation (6) is slightly modified accordingly to include the partial observability of the latent states for user i when the user responds to the survey in time period t as follows:

$$P(\boldsymbol{Y}_{i1}, \boldsymbol{Y}_{i2}, \ \ldots, \boldsymbol{Y}_{iT}|\boldsymbol{\alpha}_i, \boldsymbol{\theta}) = \boldsymbol{\pi}_i M_{i1} Q_i M_{i2} Q_i \ldots Q_i M_{it-1} Q^t_{i,\cdot\to s} M_{it} Q_i M_{it+1} Q_i \ldots Q_i M_{iT} \boldsymbol{\iota}, \quad (8)$$

which can be further modified if the researcher observes for user i the true state in multiple time periods. It is not necessary to explicitly constrain the outgoing transition matrix in Equation (8), as the outgoing paths are fully determined by the incoming paths. The likelihood function in Equation (8) is a fairly simple modification of the traditional HMM and constitutes a type of a PHMM in which the researcher observes the latent state in some but not all time periods. This PHMM may be seen as a constraint version of an HMM in which certain elements in the transition probability matrix are fixed to zero at certain time periods (Monaco and Tappert 2018). As with any constrained model, we do not expect the fit of the model to improve; however, fusing the observed survey into the model helps with calibrating the latent job-seeking states and grounding the meaning of the states. This is particularly important for applications in which state recovery, as opposed to outcome predictions, is the main objective of the modeling effort.

4.1.5. Fusion of Survey Responses with Response Error (*Stochastic* Fusion PHMM).

The proposed PHMM assumes that the survey responses fully reveal the respondent's job-seeking status and fully inform the HMM states, although it allows for one-to-many mappings between the survey responses and states (e.g., Equation (7)). However, rich literature in marketing and psychology point to the possibility that respondents may not fully reveal their true preferences or behavior in surveys (Hippler and Schwarz 1987, Schwarz 1999, Tourangeau et al. 2000). To account for possible response errors, we propose a modification to the way we fuse the survey responses to the states by allowing for a lower likelihood of moving to states that do not match the survey response at the time of the survey. This is different from the previous deterministic fusion approach, which fixes the likelihood of moving to states that do not match the survey response to 0. Specifically, $Q^t_{i,\cdot\to s}$ in Equation (7) is modified to allow for survey response errors by including in the transition probabilities a parameter γ_c, where $c = 1,2,\ldots,C$ are the survey response categories, such that the likelihood of state transitions not corresponding to the survey response of user i in period t are adjusted downward, depending on the strength of the correspondence between the survey responses and the latent states. Thus, we allow each survey response category to have a different response error. More specifically, for the month of the survey where user i responds category c to the job-seeking question, we modify each row in the transition probabilities matrix in Equation (5) as follows:

$$q^t_{ikj} = P\left(S_{it} = j | S_{it-1} = k, \ A^Q_i, \boldsymbol{\theta}\right) = \frac{\exp(\phi_{kj} + a^Q_{ikj} - \gamma_c \mathrm{I}(j \notin c))}{\sum_{l=1}^{l=K-1} \exp\left(\phi_{kl} + a^Q_{ikl} - \gamma_c \mathrm{I}(l \notin c)\right) + \exp(-\gamma_c \mathrm{I}(K \notin c))}, \quad (9)$$

for $j = 1,2,\ldots,K-1$, $k = 1,2,\ldots,K$, and $q_{ikK} = 1 - \Sigma_j q_{ikj}$. Here $\mathrm{I}(j \notin c)$ is an indicator function that equals one if state j does not correspond to survey response category c as defined by the survey to state mapping and equals 0 otherwise. When $\gamma_c = 0$, this model is the unconstraint HMM. As γ_c becomes large, the model forces the transition into the state that corresponds to the survey response.[10]

Looking at Equation (9), and following the previous example with the six-state PHMM, if user i indicates in period t that she is an active job seeker ($c = 5$), then the probabilities of transitioning into states $s = 1,2,3,$ and 4 are being pushed down depending on the magnitude of γ_5. In the extreme, if states 5 and 6 perfectly capture active job seekers ($c = 5$), then $\gamma_5 \gg 0$ and the probabilities of transitioning into states $s = 1,2,3,$ and 4 will be 0.

4.2. Model Estimation Approach

We use a Bayesian framework to estimate the PHMM, which incorporates cross-user heterogeneity to model multivariate user activity (Ebbes et al. 2010). We use a Markov chain Monte Carlo (MCMC) algorithm to directly sample the posterior distribution through Metropolis-Hastings (MH) steps (Chib and Greenberg 1995) using an adaptive tuning of the MH step (Atchadé and Rosenthal 2005). See Online Appendix B further details of the MCMC algorithm.

5. Empirical Application

We calibrate the two PHMMs described in Section 4 on the activity and survey data described in Section 3. We

fuse the responses to the job-seeking question of the survey (month 5 of the data window) into the PHMMs and use the responses to the validation survey in month 14 for holdout prediction. Of the 2,814 users who responded to the first survey, 491 users also responded to the second survey. Hence, we continue our analyses with $N = 491$ users, from whom we have validation survey responses, to examine the out-of-sample time period predictions. Furthermore, in order to predict the job-seeking status for out-of-sample users, we randomly split the data into a calibration sample ($N_{calibration} = 400$) and a validation sample ($N_{validation} = 91$).

5.1. Selecting the Number of States and the Mapping Between States and Survey Responses

For the deterministic fusion PHMM with a one-to-one mapping between states and survey response categories, the number of states is equal to the number of survey response categories (five states in our case). However, we would like to allow for the possibility that several PHMM activity states correspond to the same job-seeking status. In that case, one needs to select an appropriate mapping between the survey response category and the PHMM states. For the six-state deterministic and stochastic fusion PHMMs, there are five different possible mappings: {1,1,2,3,4,5}, {1,2,2,3,4,5}, {1,2,3,3,4,5}, {1,2,3,4,4,5}, {1,2,3,4,5,5}. For example, {1,1,2,3,4,5} corresponds to a PHMM in which survey response category 1 (non–job seeking) corresponds to two latent states and the four other response categories each correspond to one state.

Our task now is to select the number of states and to select which of the one-to-many mappings best fit the data. Bayesian model fit criteria such as the log marginal density (LMD) tend to underpenalize complex models and hence are often inappropriate for model selection in HMMs (Netzer et al. 2017). Furthermore, our interest is not to predict the outcome variables (user activity on the social network platform) but rather to predict the latent behavior, that is, the state of job search for each user. Accordingly, for model selection we use a cross-validation approach by comparing the candidate PHMM models on their ability to predict the job-seeking status in a held-out subset of the calibration sample. Specifically, we split our calibration sample ($N_{calibration} = 400$) into a training ($N_{training} = 300$) and a test sample ($N_{test} = 100$). We fit the model on data from the 300 training users and predict job-seeking status in month 5 for the remaining 100 test users.

As the number of states grows, the combinatorics of the number of possible mappings increases substantially. Indeed, in this application, given our sample size, as we attempted to increase the number of states beyond six the estimation became less stable. Hence, we restricted our analysis to up to up to 6 states. From a computational point of view, however, one can use a cloud parallel computing approach (e.g., Amazon Web Services), which allowed us to run the MCMC chains for the different versions of the PHMM in parallel to test the number of states and mapping.

In calculating the performance of different models, we compare the observed job search status from the survey with the predicted job-seeking state from the model. We use the filtering approach in each step of the MCMC sampler (Netzer et al. 2017, p. 419) to calculate the probability that user i is in state S for $t = 5$ (the survey period). A challenge arises for computing posterior state membership probabilities for the test users ($N_{test} = 100$), because we do not have estimates for the individual-level parameters ($\boldsymbol{\alpha}_i$). We therefore use the following procedure. Taking $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}$ fixed at the posterior mean estimate from the training sample, we run the observed activity in the first 4 months of the data of each test user through the MCMC sampler to generate a posterior sample of size L of random intercepts $\boldsymbol{\alpha}_i^l$, $i = 1, 2, \ldots, N_{test}$, $l = 1, 2, \ldots, L$, after which we use the filtering approach to compute $P(S_{i5}|\boldsymbol{\alpha}_i^l, \bar{\boldsymbol{\theta}}, \boldsymbol{Y}_{i1}, \boldsymbol{Y}_{i2}, \boldsymbol{Y}_{i3}, \boldsymbol{Y}_{i4})$ for the test users.

Our prediction involves a multilabel classifier among highly imbalanced classes as over 40% of the respondents responded that they would discuss a job opportunity (category 2 in Table 1). We use four commonly used prediction metrics from the statistics and machine learning literature to compare the models' performance in the cross-validation task while accounting for the imbalanced data:

1. *F1 measure*: the harmonic mean of the precision and recall.
2. *Area under the curve* (*AUC*): the area under the receiver operating characteristics (ROC) curve.
3. *Precision-recall curve* (*PRC*): because the AUC can be sensitive to imbalanced classes we also calculate the area under the *PRC*, which better handles imbalanced data (Saito and Rehmsmeier 2015).
4. *Jaccard Index (JI):* to examine the models' ability to positively predict job-seeking statuses we use the Jaccard Index, which calculates the ratio of the intersection of the true and predicted positive outcomes divided by the union of the two. This measure ignores the true negatives.

To compute the multiclass AUC, we follow the procedure outlined in Hand and Till (2001). In order to calculate the F1 multiclass score, we first compute the precision and recall scores for each of the five classes, take their average, and then compute the F1 score as the geometric mean of the averaged precision and recall. Following a similar procedure, we calculate the multiclass version of the Jaccard index. To calculate the multiclass PRC measure, we compute the PRC for each

Table 2. Prediction Performance for Cross-Validation Comparing Various PHMMs

		JI	F1	PRC	AUC
Deterministic fusion PHMM	12345	0.16	0.28	0.25	0.59
	112345	0.19	0.32	0.25	0.57
	122345	0.18	0.30	0.26	0.56
	123345	0.14	0.24	0.25	0.56
	123445	0.16	0.27	0.25	0.56
	123455	0.12	0.21	0.25	0.55
Stochastic fusion PHMM	12345	0.16	0.28	0.28	0.59
	112345	0.17	0.30	0.26	0.55
	122345	0.17	0.29	0.23	0.53
	123345	0.17	0.29	0.28	0.58
	123445	0.11	0.20	0.24	0.55
	123455	0.18	0.30	0.28	0.59

class and then take the average across the five PRC scores.

The top part of Table 2 compares the deterministic fusion PHMM with five states one-to-one mapping between the PHMM states and the survey response categories with the five six-state deterministic fusion PHMMs with the one-to-many mappings between survey responses and PHMM states. Overall, the model that splits the non–job-seeking state (response category 1) into two job-seeking states (deterministic PHMM (112345)) seems to fit the data best. The bottom part of Table 2 compares the stochastic fusion PHMM with five states one-to-one mapping between the PHMM states and the survey response categories with the five six-state stochastic fusion PHMMs with the one-to-many mappings. In this case, the model that splits the active job-seeking state into two states (stochastic PHMM (123455)) performs the best. The overall performance of the best performing deterministic and stochastic PHMMs are quite similar. Hence, we evaluate these two models further in terms of interpretation and predive ability.

5.2. PHMM Posterior Estimates

Table 3, panels A and B, reports the posterior mean and posterior standard deviation of the parameters of the three components (π, Q, and M) of the six-state determinist fusion PHMM (112345), and stochastic fusion PHMM (123455), respectively. For ease of interpretation we transformed the working parameters (α_i and θ) into posterior probabilities for the discrete variables in M, into the initial state probabilities and the transition probability matrix, and into the antilog of the expected values for the continuous variables in M. The trend parameters are reported at the working parameter level.[11]

There are several important observations to note from the posterior results in Table 3, panels A and B. First, the estimates of both the deterministic and stochastic fusion models are consistent with the model-free evidence (Section 3.3). That is, active and casual job seekers (states 5 and 6 in Table 3, panel A and states 4–6 in Table 3, panel B) are more likely to update their profile, search for jobs, search on the platform for other information than jobs, and visit more pages. In terms of social activity, those who actively search for a job, tend to send more invitations to connections outside their current company, and they tend to send more invitations than they receive (the ratio is 12.88/3.08 = 4.18 in Table 3, panel A, and 52.65/5.16 = 10.20 in Table 3, panel B), compared with the others for whom this ratio is more balanced. Additionally, active job seekers tend to form more connections that are well connected themselves. This finding suggests that there is some strategic networking behavior among job seekers on the platform.

Second, we observe an interesting pattern when we consider the two non–job-seeking states in the deterministic PHMM (Table 3, panel A). There are non–job seekers (state 1) that use the platform very little, with few profile updates, searches, and pageviews. Users in this state are not actively growing their network either. At the same time, there is a second group of non–job seekers (state 2) that use the platform quite frequently, approximately at the level of casual job seekers. However, these active non–job seekers use the job search tool considerably less often than the casual job seekers. Thus, they are active on the platform but not in a job-seeking manner. At the same time, users in this state have a fairly high probability of transitioning into the *thinking about job search* state and perhaps start exploring the platform vehicle for job search.

Similarly, the stochastic PHMM (Table 3, panel B) that splits active job seekers into two states finds a job-seeking state that is very active on the platform with respect to almost every activity (state 6) and a job-seeking state in which users only use the platform moderately (state 5), less, on average, than users in the casual job seekers state (state 4). Whereas the active job seekers who are using the platform frequently are most likely to transition to state 5 (active job seekers who use the platform less frequently), active job seekers who use the platform less frequently are most likely to transition into a passive job-seeking state (state 3). This may signal different stages in the job search process, wherein job seekers are moving from an active exploratory search using the social network platform to an offline more targeted search at a few companies.

Third, considering the transition probability matrix, it is reassuring to observe that the diagonal elements are the highest in most rows, suggesting that users are more likely to stay in their job-seeking state from one month to another. Specifically, the low activity non–job-seeking search states are most sticky. Based on the deterministic model with only one job-seekers state (state 6 in Table 3, panel A), if a user is in the active job-seeking state in month t, then the user's probability of

Table 3. Posterior Means (Standard Deviations)

Panel A: Deterministic PHMM (112345)														
State	1		2		3		4		5		6			
Survey response	1		1		2		3		4		5		Trend	
Profile updates (dum)	0.02	(0.01)	0.38	(0.02)	0.19	(0.02)	0.09	(0.02)	0.32	(0.03)	0.55	(0.02)	−0.01	(0.00)
Job searched (dum)	0.00	(0.00)	0.14	(0.03)	0.03	(0.01)	0.07	(0.01)	0.73	(0.03)	0.62	(0.03)	0.01	(0.01)
Total searches	1.76	(0.50)	11.56	(0.95)	4.61	(0.43)	2.79	(0.22)	7.70	(0.77)	34.83	(2.90)	0.01	(0.00)
Pageviews	7.81	(0.59)	142.19	(6.23)	63.65	(4.37)	52.58	(2.63)	125.57	(7.31)	356.36	(16.45)	0.02	(0.00)
More invitations outside company (dum)	0.32	(0.29)	0.85	(0.04)	0.86	(0.05)	0.65	(0.11)	0.81	(0.12)	0.90	(0.03)	0.04	(0.05)
Invitations sent	0.06	(0.10)	4.17	(0.33)	3.98	(0.55)	2.73	(0.40)	2.69	(0.34)	12.88	(1.28)	0.01	(0.01)
Invitations received	1.80	(0.24)	3.72	(0.23)	2.31	(0.24)	2.12	(0.09)	2.26	(0.32)	3.08	(0.28)	0.00	(0.01)
Connections formed	1.50	(0.25)	7.81	(0.42)	3.14	(0.22)	2.88	(0.14)	3.07	(0.33)	12.04	(0.97)	0.01	(0.01)
Log number of connections of invitee	6.21	(4.35)	5.68	(0.27)	5.32	(0.33)	5.24	(0.50)	5.04	(0.68)	7.54	(0.23)	0.03	(0.01)
Initial state distribution	0.37	(0.02)	0.14	(0.02)	0.11	(0.03)	0.20	(0.02)	0.10	(0.03)	0.08	(0.02)		
Transition matrix														
From 1 to …	0.42	(0.03)	0.10	(0.02)	0.23	(0.03)	0.13	(0.02)	0.07	(0.02)	0.05	(0.01)		
From 2 to …	0.08	(0.03)	0.34	(0.03)	0.11	(0.03)	0.40	(0.04)	0.01	(0.01)	0.06	(0.01)		
From 3 to …	0.16	(0.02)	0.17	(0.03)	0.35	(0.05)	0.19	(0.04)	0.07	(0.02)	0.05	(0.01)		
From 4 to …	0.17	(0.02)	0.15	(0.02)	0.22	(0.03)	0.35	(0.04)	0.07	(0.02)	0.04	(0.01)		
From 5 to …	0.26	(0.05)	0.04	(0.03)	0.07	(0.03)	0.19	(0.05)	0.36	(0.04)	0.08	(0.02)		
From 6 to …	0.09	(0.02)	0.12	(0.03)	0.11	(0.03)	0.12	(0.03)	0.06	(0.02)	0.50	(0.03)		
Panel B: Stochastic PHMM (123455)														
State	1		2		3		4		5		6			
Survey response	1		2		3		4		5		5		Trend	
Profile updates (dum)	0.02	(0.00)	0.25	(0.03)	0.08	(0.01)	0.51	(0.03)	0.29	(0.02)	0.62	(0.03)	−0.01	(0.00)
Job searched (dum)	0.00	(0.00)	0.16	(0.02)	0.02	(0.01)	0.49	(0.03)	0.32	(0.02)	0.72	(0.03)	0.00	(0.00)
Total searches	1.12	(0.54)	3.67	(0.29)	1.69	(0.19)	26.57	(2.08)	7.60	(0.58)	62.37	(8.27)	0.00	(0.00)
Pageviews	5.65	(0.42)	55.22	(3.15)	31.13	(1.52)	277.17	(11.62)	116.77	(4.68)	719.36	(46.98)	0.01	(0.00)
More invitations outside company (dum)	0.20	(0.25)	0.96	(0.06)	0.33	(0.18)	0.84	(0.03)	0.87	(0.03)	0.93	(0.03)	0.01	(0.04)
Invitations sent	0.05	(0.08)	2.55	(0.40)	1.58	(0.26)	12.56	(0.68)	2.85	(0.17)	52.65	(5.81)	0.00	(0.01)
Invitations received	1.80	(0.23)	1.65	(0.33)	2.13	(0.07)	3.02	(0.23)	3.03	(0.15)	5.16	(0.66)	0.01	(0.00)
Connections formed	1.57	(0.25)	1.92	(0.20)	2.51	(0.08)	13.20	(0.54)	5.66	(0.22)	43.87	(3.01)	0.00	(0.00)
Log number of connections of invitee	4.70	(4.51)	4.90	(0.63)	4.68	(0.59)	7.61	(0.12)	5.08	(0.24)	9.91	(0.16)	0.03	(0.01)
Initial state distribution	0.36	(0.02)	0.17	(0.02)	0.17	(0.02)	0.06	(0.01)	0.23	(0.02)	0.01	(0.01)		
Transition matrix														
From 1 to …	0.46	(0.03)	0.13	(0.02)	0.17	(0.02)	0.07	(0.02)	0.15	(0.03)	0.02	(0.01)		
From 2 to …	0.20	(0.02)	0.36	(0.04)	0.18	(0.03)	0.05	(0.01)	0.21	(0.03)	0.00	(0.00)		
From 3 to …	0.15	(0.01)	0.15	(0.03)	0.35	(0.02)	0.05	(0.01)	0.30	(0.03)	0.01	(0.00)		
From 4 to …	0.10	(0.03)	0.03	(0.03)	0.18	(0.04)	0.27	(0.03)	0.39	(0.04)	0.03	(0.01)		
From 5 to …	0.11	(0.02)	0.15	(0.03)	0.27	(0.03)	0.09	(0.01)	0.35	(0.03)	0.03	(0.01)		
From 6 to …	0.00	(0.00)	0.01	(0.01)	0.01	(0.01)	0.11	(0.03)	0.14	(0.04)	0.73	(0.04)		

being again in the active job-seeking state in the next time period is 0.50, which corresponds to a duration of about two months of active job seeking. This result is fairly consistent with the reported median duration of unemployment of approximately 10 weeks).[12]

Table 4 presents the posterior results for the penalization parameters (γ_c's in Equation (9)) for the stochastic fusion PHMM. Recall that higher values reflect a stronger correspondance between the survey response and the PHMM states. We can see that the casual and active job-seeking response category have the highest survey response correspondance.

5.3. Posterior Predictions of Job Search Status

To identify the job-seeking status of its users, the platform needs to predict the job-seeking status of the entire user base over time, as it is impossible to survey all users in every time period. Thus, the company needs to predict the job-seeking status of users who never responded to a job-seeking survey and the status of users who responded to a survey in one time period for the remaining time periods. To test the model for such prediction scenarios, we consider (a) predicting the survey response of out-of-sample users ($N_v = 91$), who were not used for model calibration, and (b)

Table 4. Posterior Means (Standard Deviations) for the Penalization Parameters (γ_c's) in the Stochastic Fusion PHMM

Survey response category	Posterior mean	Posterior standard deviation
1	0.62	(0.24)
2	0.56	(0.20)
3	0.57	(0.24)
4	0.70	(0.29)
5	0.68	(0.29)

predicting users in out-of-sample time periods, that is, predicting the users' response to the validation survey, which occurred approximately one month after the end of the calibration data window. Table 5 summarizes our prediction schema for out-of-sample periods and users. We note that, unlike other applications of HMMs in marketing, our objective is not to predict the state-dependent activities (M) in future periods but rather to predict the latent states of the users.

We consider three types of holdout predictions (Table 5):

1. For the calibration sample ($N_c = 400$), we predict the job-seeking status in month 14. These predictions test the model's ability to predict the job-seeking status for users who were previously surveyed by the firm but who's current job-seeking status is unknown.

2. For the holdout sample ($N_v = 91$), we predict the job-seeking status in month 5. These predictions test the model's ability to predict the job-seeking status for users who were never surveyed but for a time period in which some users were surveyed. We use only the observed activity during the first four months of the holdout sample to predict the job-seeking status of these users in month 5.

3. For the holdout sample ($N_v = 91$), we predict the job-seeking status in month 14. This represents the most challenging prediction scenario to test our model: predicting for users who were not surveyed before during a time period in which no survey was conducted. Arguably, this scenario reflects the most typical business case, as survey sample sizes generally are small relative to the total userbase (millions of users in our case). Hence, this scenario is the cleanest and most practical prediction scenario to test our model.

We do not predict the job-seeking status in month 5 for users who responded to the survey as these data were directly fused into the PHMMs. We compare the predictions of four versions of PHMMs with the prediction ability of two machine learning benchmark models that use all the available user activity in the four months before the prediction month. Machine learning models, and specifically the RF model, have been used by the company we collaborated with to identify potential job seekers. Our set of models is as follows:

1. *Det. PHMM 12345*: The deterministic fusion five-state PHMM with one-to-one mapping between survey responses and job-seeking states.

2. *Det. PHMM 112345*: The deterministic fusion six-state PHMM with two HMM states for the non–job-seeking status.

3. *Stoch. PHMM 12345*: The stochastic fusion five-state PHMM with one-to-one mapping between survey responses and job-seeking states.

4. *Stoch. PHMM 123455*: The stochastic fusion six-state PHMM with two HMM states for the active job-seeking status.

5. *Lasso:* A regularized ordered logit regression with the survey responses as the dependent variable and lagged user activity in the four months before the job-seeking prediction period as predictors.

6. *RF*: Similar to the Lasso regression but with a RF ordered logit model that allows for possible nonlinearities in the relationship between platform activity and job-seeking behavior.

To calibrate the Lasso and RF models, we regress the observed survey response in month 5 as an ordinal variable on the same (nine) variables that were used to calibrate the PHMMs in months 1, 2, 3, and 4. To predict the job-seeking status for the second survey we use the user activity in months 11,12, 13, and 14. Thus, the machine learning models include dynamics via the lagged observed activities as covariates, making them strong contenders to the PHMMs as they fit

Table 5. Schematic Overview of the Prediction Analyses

	Time	
Cross section	Month 5: Survey	Month 14: Validation survey
Calibration sample ($N_c = 400$)	In-sample users and in-time period No predictions are made as the users' job-seeking status was directly fused into the PHMMs for these users in that time period.	[1] In sample users, out-of-time period Predict job-seeking status in month 14 for users whose responses to the survey were used to calibrate the model.
Holdout sample ($N_v = 91$)	[2] Out-of-sample users and in-time period Predict job-seeking status in month 5 for a hold-out sample of users at the time period of the survey.	[3] Out-of-sample-users, out-of-time period Predict job-seeking status in month 14 for a hold-out sample of users at a time period after the calibration time period.

directly the variable of interest, job-seeking status, as a function of past activity.[13]

We compare the six models with respect to the same prediction metrics used to choose the number of states and the mapping between states and the survey job-seeking categories in Section 5.1. We use the same procedure described in Section 5.1 to obtain the individual level parameters ($\boldsymbol{\alpha}_i$) for the holdout sample users ($N_v = 91$). Table 6 shows the prediction results for the six models.

First, we observe from Table 6 that the PHMMs generally predict job-seeking status better than the two machine learning benchmark models. Among the two machine learning approaches, the RF predicts better than the Lasso. The performance of the four PHMMs is quite similar. The RF approach is comparable to the PHMMs on some measures, but it falls short on other measures and in particular on measures that emphasize recall (identifying job seekers) such as that Jaccard Index.

The prediction analysis in Table 6 includes only one split of the data into a calibration and a validation set, which, while random, may be a particular allocation of observations to the calibration and validation sets. In Online Appendix E, we present a robustness analysis of these predictions using a fivefold cross validation approach. Our results and conclusions remain the same.

We next explore how the proposed PHMM and the RF approaches perform in predicting job-seeking *duration*.

5.4. Capturing the Duration of Active Job Search

Thus far, we focused on predicting the job-seeking status of a user in a particular month and demonstrated the potential benefit of the proposed PHMMs over machine learning–based benchmark models. In this section, we demonstrate that this potential benefit arises from the PHMM's ability to better capture dynamics in job-seeking status. By the nature of the latent states, and the transitions among them, the PHMM should be able to capture how long a user has been actively searching for a job and when a user transitions into the active job-seeking state. This is important for the platform, because it is particularly interested in identifying those users who are starting to actively search for a job and knowing when a job seeker stopped seeking for a job. In order to test the ability of the proposed model to capture job search duration, we asked respondents in the validation survey *how long* they had been job seeking. We emphasize that the validation survey was not used for calibrating the PHMM. We use the deterministic fusion PHMM 112345 to predict the job-seeking status of the user in months 8–14.[14] We then split the users by their validation survey response into two groups: those who were actively searching and those who were not searching for a job. Respondents who indicated they were actively job searching, were further split into two groups of job search duration: (1) those who were actively searching for at most three months, and (2) those who were actively searching for more than three months.

If the PHMM predicts job searching well, we should see that those who are actively searching for a job according to their response to the validation survey have a higher likelihood of being in the active job-seeking state in month 14, relative to users who are not seeking for a job. Moreover, we should find that users who indicate in the validation survey that they have been actively searching for a job for at most three months, transition from a low probability of being in the active job-seeking state up to month 11 to a higher probability of being in the active job-seeking state after month 11.

To provide some model-free evidence, we investigate the average differences in the four user activities (i.e., page views, total searches, job searches, and

Table 6. Holdout Predictions for the Three Versions of the PHMM and the RF and Lasso Ordered Logit Benchmarks

		Month 5: Survey				Month 14: Validation survey			
Cross section	Model	JI	F1	PRC	AUC	JI	F1	PRC	AUC
Calibration sample ($N_c = 400$)	Deterministic PHMM 12345					0.14	0.24	0.23	0.55
	Deterministic PHMM 112345					0.14	0.24	0.24	0.55
	Stochastic PHMM 12345		N/A			0.12	0.21	0.22	0.53
	Stochastic PHMM 123455					0.13	0.24	0.23	0.53
	RF					0.11	0.19	0.22	0.54
	Lasso					0.08	0.16	0.21	0.51
Holdout sample ($N_h = 91$)	Deterministic PHMM 12345	0.16	0.28	0.27	0.59	0.21	0.34	0.29	0.60
	Deterministic PHMM 112345	0.13	0.22	0.27	0.60	0.19	0.32	0.31	0.60
	Stochastic PHMM 12345	0.15	0.26	0.26	0.55	0.17	0.29	0.30	0.57
	Stochastic PHMM 123455	0.16	0.28	0.27	0.57	0.18	0.30	0.25	0.55
	RF	0.11	0.19	0.26	0.61	0.10	0.19	0.26	0.58
	Lasso	0.13	0.23	0.24	0.55	0.08	0.16	0.24	0.54

Notes. Performance metrics (JI, F1, PRC, AUC) indicate model performance to predict the users' job-seeking status in month 5 and month 14. Higher numbers indicate better performance. See definitions of JI, F1, PRC and AUC in Section 5.1.

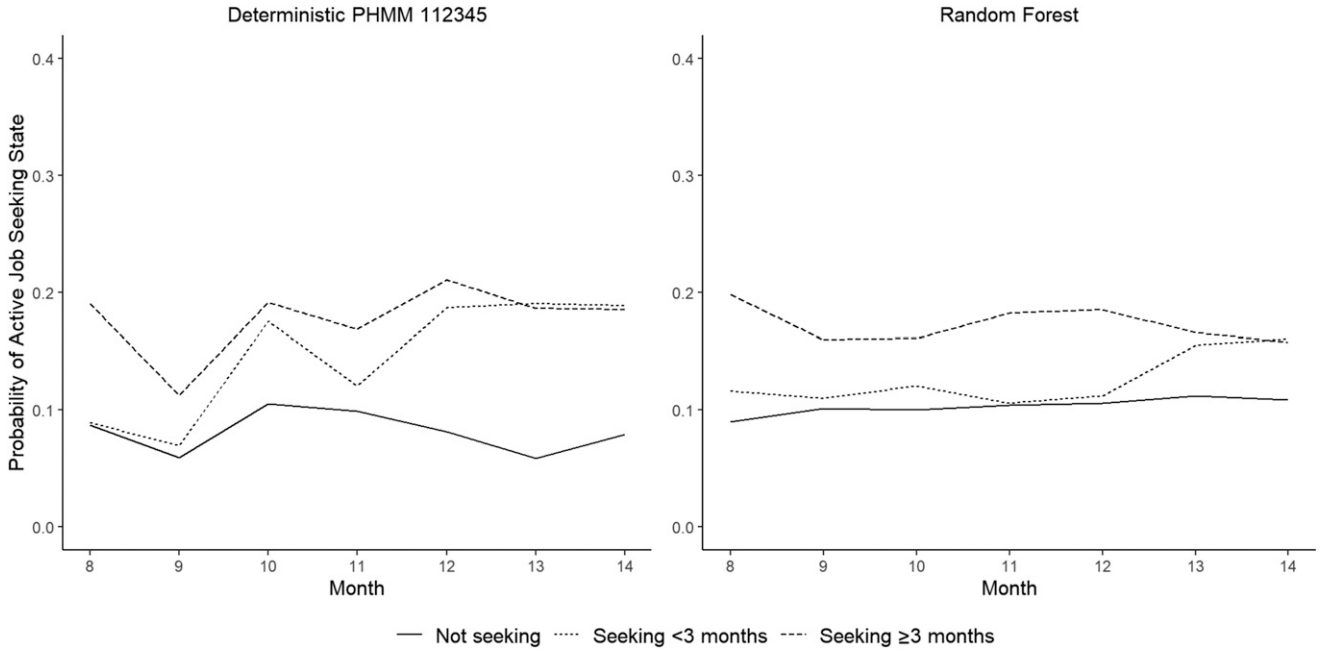
profile updates) that we observe before the validation survey between the job-seeking periods and the non–job-seeking periods for the users who indicate in the validation survey that they have been actively searching for a job for up to three months. We find a significant difference for page views (on the log-scale; $p < 0.001$) and total searches (on the log-scale; $p = 0.011$), a marginal difference for job searches ($p = 0.093$) and insignificant difference for profile updates ($p = 0.447$). The insignificant difference for profile updates may stem from the fact that profile updates are often done earlier on in the job search process. This analysis suggests that the differences in behavior between job seekers and nonjob seekers are not merely because of differences between the two groups of users but because of a change in behavior of the *same* user as the user transitions to a job search status.

Moving beyond model-free evidence, the state predictions of the PHMM and the strongest contender machine learning model (the RF model) are provided in Figure 2. Several interesting observations can be made from Figure 2. Consistent with the results in Table 6, the PHMM separates job seekers from non–job seekers in month 14 (the approximate time of the validation survey). That is, the likelihood of being in the active job-seeking state in month 14 is considerably higher for those who report being job seekers (dotted and dashed lines) than for those who report not being job seekers (solid line). The separation in month 14 is less strong for the RF model, indicating that this model does not do as well in separating job seekers from non–job seekers.

More importantly, comparing the dashed and dotted lines, we see that the PHMM does well in, not only predicting who is job seeking, but also in predicting *when* the users transitioned to the active job-seeking state. Specifically, for those users who indicate that they were active job seekers for at most three months (dotted line), the PHMM shows a transition from a behavior similar to non–job seekers before month 11, to a behavior consistent with active job seekers after month 11. For those who state in the validation survey that they have been actively searching for a job for more than 3 months (dashed line), we see a consistently higher probability of being in the active job-seeking state relative to those who state they were not job seeking in the validation survey (solid line). Unlike the PHMM, the RF model is not able to pick up this signal well.

We quantify the effects in Figure 2 for users who transitioned from a non–job-seeking status to a job-seeking status by comparing the average estimated state probabilities of being an active job seeker for the months the user was job seeking to the months the user was non–job seeking according to the user's response to the duration question in the second survey (how many months have you been job seeking). For the PHMM, the average active job seeking state probability was 0.21 during the active job-seeking months

Figure 2. Average Probabilities of Being in the Active Job-Seeking State for Months 8–14 for the PHMM (Left) and RF (Right)



Notes. Dashed line: the average probability of being in the active job-seeking state for users that indicated in the validation survey that they were actively searching for 3 months or longer. Dotted line: the average probability of users that indicated in the validation survey they were actively searching for a job for at most three months. Solid line: the average probability for users that indicated in the validation survey they were not searching for a job.

Table 7. Estimates of a Random-Effect Logit Discrete Time Hazard Model of Transitioning into Active Job-Seeking Behavior with the Predicted Active Job-Seeking State Probabilities of the PHMM and the RF Model as Predictors

	Model 1		Model 2		Model 3	
	Estimate (standard error)	p value	Estimate (standard error)	p value	Estimate (standard error)	p value
Constant	−1.01 (0.29)	0.001	−0.83 (0.34)	0.014	−0.96 (0.39)	0.013
PHMM	2.00 (0.80)	0.013			2.05 (0.84)	0.015
RF			1.23 (1.27)	0.488	−0.37 (2.02)	0.854
Deviance	224		232		224	

and 0.10 during the prior non–job-seeking months (t = 2.92, p = 0.007). This difference was much smaller and not statistically significant for the RF model (0.16 versus 0.15, t = 0.60, p = 0.551), suggesting that the RF did not detect the transition into active job-seeking behavior.

The results in Figure 2 and corresponding analysis demonstrate a potential important benefit of the proposed PHMM; it can detect changes over time in users' likelihood of being in a job-seeking state and may be used to early detect changes in the user's job-seeking status. The analysis in Figure 2 has the limitation that it discretizes the active job search duration into three months or less and more than three months. However, the survey respondents in the validation survey provided a continuous measure of how many months the user has been job searching. To leverage that information, we estimate a random-effects logit discrete time proportional hazard model (Gupta 1991) that predicts the hazard of transitioning into active job-seeking behavior (where 1 = job seeking and 0 = not job seeking in the focal month), with the PHMM predicted active job-seeking state probabilities, and the RF-predicted active job-seeking status probabilities, in months 8–14, as predictors.

Table 7 shows that the deterministic fusion PHMM active job search state probabilities (state 6) positively and significantly predict the hazard of transitioning into an active job-seeking status (model 1). In contrast, the RF active job-seeking probabilities do not significantly correlate with the hazard of transitioning into an active job-seeking status (model 2). When including both probabilities (PHMM and RF model) as predictors in the same hazard model (model 3), only the PHMM probabilities significantly correlate with the hazard of active job seeking. Additionally, the model fit (as measured by the deviance) of model 1 is similar to that of model 3, suggesting that once the PHMM state probabilities are included, the RF probabilities do not add additional information.

In sum, our findings in this section provide convergent support for the proposed model as an approach to infer and predict latent behavior (job-seeking status) from observed activity that are indirect noisy proxies of the latent behavior of interest. Our proposed model not only provides rich insights into job-seeking behavior but it also predicts job-seeking status better than machine learning-based approaches, because it more effectively captures transitions between active and nonactive job-seeking behaviors. Such information may be used for targeting purpose as we demonstrate next.

6. Targeting Job Seekers

From a marketing perspective, the social network platform is interested in detecting job seekers in order to target such users with relevant marketing offers. We demonstrate how the proposed approach can be used to profitably target potential job seekers through the platform's internal direct mail tool (for convenience we will abbreviate this tool as d-mails). d-mails are among the most common recruiting tools on the platform. d-mails serve as an internal cold-call tool allowing strangers on the social network platform to email users they are not connected to. According to the platform, this tool is often used by recruiters to identify potential candidates. Thus, the effectiveness of a d-mail should increase if it is being sent to a job seeker instead of a non–job seeker. At the time of the data collection, a d-mail cost $10 per d-mail if the user responded to the d-mail within 7 days. If the user did not respond to the d-mail within 7 days, the sender would receive a $10 credit back. In other words, from a profitability point of view, it is important for the platform that users respond to d-mails. We examine whether targeting d-mails to those users who are identified by our proposed approach as job seekers would lead to higher response rates and higher profits.

The data (Section 3) used to calibrate and validate the model (Sections 4 and 5) did not include exposure and responses to d-mails. However, we obtained from the data provider a second user activity data set that includes, in addition to the user activity on the platform, information on whether and when the user received a d-mail and whether the user responded to it. This second data set also allows us to test our PHMM

approach to model job-seeking behavior for a new set of users. The second data set includes 1,621 users for whom we observe their activity on the platform during the 12-month time period of June 2011–May 2012. As before, we observe a response from these users to a job search survey in the fifth month of the data window.[15] We estimate the deterministic fusion PHMM 112345 for this second data set. The interpretation of the states and model estimates are consistent with those found for the first data set in Section 5 (See Online Appendix F for the posterior estimates of the parameters of the PHMM for this data set).

For the set of users observed in this sample, we observe whether and when they received a d-mail, and, if they received a d-mail, whether they responded to it. Overall, during the 12-month data period, 864 d-mails were sent, an average of 0.53 d-mails per user, with 317 positive responses. First, we examine the 72 d-mails (and 21 positive responses) that were sent during the month of the survey to the 1,621 users (Table 8). Because of the relatively small number of d-mails with positive responses, we analyze the data at three (rather than five) job-seeking response categories. Following the platforms' classification of the response categories, we define non–job seekers as those who responded in the survey that they are not job seeking (response category 1), passive job seekers as those who responded that they would discuss or are thinking about job seeking (response categories 2 and 3), and active job seekers as those who responded that they are casually or actively job seeking (response categories 4 and 5). We note that the platform used this classification of the survey responses into three categories when analyzing the survey results. Table 8 shows that the d-mails were sent with approximately equal probability to the three job-seeking status types. However, active job seekers are more likely to respond to d-mails (33.3%) than non-job seekers (14.3%). That is, senders of the d-mails do not seem to identify and/or consider the job-seeking status of the users, despite the potential higher response rate of active (and passive) job seekers. One possible explanation is that senders have no obvious way of recognizing who is an active job seeker on the platform. Thus, there may be an opportunity to improve the effectiveness of d-mails by targeting users based on their inferred job-seeking status. This is of particular financial importance to the platform because it does not collect any revenue for d-mails to which users did not respond. Accordingly, we compare the current policy of sending d-mails with a policy that prioritize sending d-mails to those who are identified as job seekers based on our proposed model.

We consider the 864 d-mails sent during our period of observation for which we observe the users' actual response. We evaluate a policy that sends 100 d-mails and targets users based on the following:

Table 8. d-Mails Received and Responded to in the Month of the Job-Seeking Survey Based on the Users' Responses to the Survey

Job seeking state (response to survey)	Probability of receiving d-mails	Probability of response to d-mails (given received)
Non-job seeker	3.3%	14.3%
Passive job seeker	5.0%	32.5%
Active job seeker	4.6%	33.3%
N	1,621	72

1. *Current policy*, for which we select 100 d-mails randomly from the set of 864 d-mails observed in our data. This policy mimics the policy observed in the data.
2. *A job-seeking state policy*, for which we rank the 864 users who received a d-mail based on their predicted probability of being in the job-seeking states (states 5 and 6) according to the proposed deterministic fusion six-state PHMM (112345), and subsequently select the 100 users with highest probabilities as targets.

We evaluate the policies based on the actual responses from the targeted users. The current policy results in a 36.5% response rate, leading to a profit for the platform of \$3.65 per d-mail sent. On the other hand, when the same 100 d-mails are targeted to those with the highest likelihood of being in the job-seeking states of the PHMM, the response rate increases to 47%, resulting in a profit for the platform of \$4.7 per d-mail sent. This corresponds to a 29% lift in profit. Given the number of d-mails sent on the platform every month, such a lift in profit could have substantial financial implications.

7. Conclusion

Many companies nowadays observe rich customer activity data that they can use for targeting customers. However, consumers' motivation and hence the basis for targeting are often not driven by customers' observed traits but rather by their latent states such as job seeking, expecting a child or a relocation. In order to successfully target customers, it is, therefore, important to identify the customers' latent states from their observed behavior. The targeting of customers may be particularly important for the firm during periods of transition from one state to another in order to make appropriate and timely offers to the customer.

We develop two versions of a PHMM to uncover the latent states of job search using data from an online social network platform with a substantial professional networking component. From a methodological point of view, unlike most marketing applications of HMMs, our research demonstrates the usefulness of HMMs to uncover and predict the latent states as opposed to predict activity given the state. Furthermore, we extend the traditional HMM framework to a

PHMM framework that naturally fuses longitudinal (social network) activity data with (possibly noisy) one-time survey data that asks users about their latent state. This is particularly useful for applications where detecting the latent state of the customer is of major business importance to the firm, as is the case for the social network platform we collaborated with.

We demonstrate that the proposed PHMMs more accurately predicts the users' job-seeking statuses, both for out-of-sample users and out-of-time periods, compared with, for instance, the RF machine learning model. Importantly, we show that the proposed approach predicts how long users are actively job searching and when they transition into the active job-seeking state, whereas the RF model was not able to capture such dynamic patterns. Additionally, our proposed approach allows the firm to identify which platform activities are associated with job search. Finally, we demonstrate the marketing value of predicting the latent states by applying the proposed approach to a targeting campaign. Using data from a past targeting campaign, we show that targeting based on the users' predicted job-seeking status from the proposed model can result in a profit lift of 29%, offering a considerable improvement over the targeting practice observed in the data.

In this paper, we obtained a rather unique and rich data set from a social networking platform about users' activity on the platform as well as their responses to two waves of a job-seeking survey. However, as with any data set, there are also limitations to our data.

First, there may be some degree of self-selection in terms of responding to the surveys by more active users on the platform. Thus, our data fusion for users who did not respond to the survey may not be MAR. To investigate the extent to which our violation of MAR is related to job-seeking behavior, we compare users who responded to both surveys (N = 491) to users who responded to only the first survey (N = 2,323). We find that those who responded to both surveys are indeed, on average, more active on the platform, by visiting more pages and conducting more searches. However, they do not update their profile more often. Importantly, comparing the survey responses of the two groups to the job-seeking question in the first survey, we find that there is no significant difference in their responses ($\chi^2(4, N = 2,814) = 2.08$, $p = 0.72$). In addition, the first survey also included a few demographic questions, such as age, gender, and income. Although 50+ users were more likely to answer both surveys, there was no difference in gender and income between those that answered the survey once or twice. Interestingly, there was also no difference in employment status (full time, part time, self-employed, unemployed, or other) at the time of the first survey between those that only answered the first survey and those that also answered the second survey ($\chi^2(4, N = 2,814) = 0.85$, $p = 0.93$). Thus, we conclude that, while our data fusion may violate MAR with respect to platform activity, it does not with respect to basic demographic characteristics, and more importantly, to our variables of interest – job seeking and employment status. We note that our results should be particularly applicable to the somewhat more active user group. It would be very difficult to identify job-seeking status (or anything for that matter) for users with very limited activity on the platform. Future research could explore ways to model the missing data mechanism if data are also available for users who were not exposed to a survey (Kamakura and Wedel 2000).

Second, one may argue that asking users about their job-seeking status may prompt users to start searching for a job and become more active on the platform. That is, a mere-measurement effect (Morwitz et al. 1993) would explain the high activity observed once the users receive the job-seeking status survey. If this were the case, then we should see an increase in activity for *all users*, including those who responded to the survey to be non-job seekers on or following the month of the survey (month 5). As shown in Figure 1, the average activity level of non–job seekers does not exhibit such an increase. Another reason why we do not believe that our results have mere-measurement effects is that the validation survey was fielded shortly after the end of the data collection period (month 14). If the results were driven primarily by mere-measurement, we would not be able to predict the job-seeking survey responses from activity before the validation survey, because by definition, mere-measurement effects can only occur after the measurement.

Third, one could argue that one may use the user's profile information, particularly position and/or company change, to identify job seeking instead of using the survey responses. Based on discussions with the data provider and preliminary data analysis, we conclude that such proxies are unreliable indicators for job-seeking status. According to the data provider, users are often unreliable in promptly updating their profile page following a successful job search. In fact, users often wait with updating their profile page until their next job search. Our data support this notion. We find that those who were job searching according to their survey response in month 5, were more likely to modify their position during the three months before the survey than those who were not job seeking ($F(1,2182) = 51.11$, $p < 0.001$). This finding suggests that position change may be an indicator of a future job search rather than an indicator of a past job search. Additionally, whereas company or position change may signal the end of a successful job search, these indicators would not identify those who have been job

searching for a long time nor those who searched for a job but decided to not take it.

Fourth, we are constrained in our analysis by the sample size of the survey responses, which is relatively small compared with the user base of most online social networking platforms. One may wonder about the scalability of our proposed approach to the typical size of the user base on the platform. We note that, although estimating the proposed approach on the sample of users is computationally intensive, our out-of-sample prediction approach is scalable. Specifically, our approach to estimate α_i for the out-of-sample users and predict their latent job search state is rather fast, can be run on parallel processors, and is therefore scalable to a large user base (Section 5.2). Similarly, as the number of states increases, we recommended using parallel processor machines or cloud computing to run the models with different matching between surveys responses and states in parallel and choose the appropriate model based on test data predictive measures (Section 5.1).

To conclude, in this research, we identify latent (job seeking) states from activity on a large social network platform. We believe that the proposed approach is applicable to many business settings where firms need to identify customers' unobserved life transitions, such as expecting a child, relocating, buying a house or going to college, from noisy observable signals. We encourage future research to explore such settings using our proposed modeling approach. We believe that our proposed approach that fuses survey responses for a sample of customers with longitudinal activity data through latent state modeling is a promising avenue to take.

Acknowledgments

The authors thank Eva Ascarza for helpful discussions and comments on this manuscript.

Endnotes

[1] At the request of the firm that provided the data, we do not disclose the company name. However, identifying who is job seeking is at the heart of the firm's business model, and job seeking is an important reason for users to engage with the social network platform. Furthermore, many recruiters use the platform to evaluate candidates. According to the firm and its financial reports, a substantial part of the firm's revenue comes from targeting job seekers.

[2] See https://www.statista.com/statistics/873648/us-staffing-industry-market-size/ (last accessed September 2020).

[3] The sample was fully anonymized (i.e., we do not observe the identity of the users or of their connections, nor do we observe the user's personal profile page). The sample was drawn from the platform's U.S. user base. We have limited information regarding the social connections of the users. At the request of the data provider, we also masked the absolute monthly activity levels by multiplying them with a random number, which was a single draw from a uniform distribution on the interval [0.5, 1.5], in all tables and figures.

[4] This variable includes any update of the profile page, such as picture, title, education, or bio. We found that updates of each aspect of the profile were too infrequent to include as separate variables in our model for this sample. Similarly, multiple profile updates per month were not frequent enough to treat this variable as a count variable in our model. Hence, we collapsed these profile update types into a single dummy variable.

[5] We choose a Tobit type 1 rather than a type 2 specification because the 0s in our data do not arise from a particular participation process (e.g., self-selection), and simply reflect lower bound activity. As we do not model the participation process, a type 1 Tobit is the natural choice.

[6] Bolding was added for exposition purposes in the paper but was not present in the actual survey. This question was designed for the data provider by an external consulting firm.

[7] At the time of our study, the U.S. unemployment rate was nearly 10%, which resembles the responses to "I am actively looking for a new job and sharing my resume," providing some face validity to these survey responses (source: Bureau of Labor Statistics 2017).

[8] We obtained the number of unique visitors to the platform in each quarter (interpolated to the monthly level) from the data provider.

[9] To allow for reliable estimation of the random-effect parameters, we do not include random-effect intercepts for the state-dependent behavior of the discrete variables and the continuous variables that we observe for only five time periods (how many new connections the user formed, how many invitations the user sent or received, and how many connections, on average, the new connections of the user had).

[10] We also estimated a different version of the stochastic PHMM in which, instead of modifying the transition matrix, we include in the state dependent activity distribution m_{itj}, another activity in the month of the survey that captures the survey response. See Online Appendix A for details. That version of the stochastic fusion PHMM performed less well than the one presented here.

[11] The posterior means and standard deviations of the working parameters are available from the authors upon request. In Online Appendix C, we discuss the posterior results for the upper-level model (heterogeneity).

[12] See https://www.bls.gov/opub/ted/2011/ted_20110602.htm (last accessed September 2020).

[13] We report details of the Lasso and RF models and results in Online Appendix D.

[14] Given that the different PHMMs perform rather similarly (Table 6), we present here the results for the deterministic fusion PHMM 112345 and compare it with the RF approach. We obtain similar results when using the stochastic fusion PHMM 123455 model.

[15] For this sample, we observe a slightly different set of activities compared with the first data set. Specifically, we observe whether the user viewed any jobs on the platform, whether the user updated the education and/or position section of the profile page, the number of invitations received and sent by the user, the number of pages the user viewed, and the number of people that viewed the user's profile page.

References

Amemiya T (1984) Tobit models: A survey. *J. Econometrics* 24(1-2):3–61.

Ansari A, Montoya R, Netzer O (2012) Dynamic learning in behavioral games: A hidden Markov mixture of experts approach. *Quant. Marketing Econom.* 10(4):475–503.

Ascarza E, Hardie BG (2013) A joint model of usage and churn in contractual settings. *Marketing Sci.* 32(4):570–590.

Ascarza E, Netzer O, Hardie BG (2018) Some customers would rather leave without saying goodbye. *Marketing Sci.* 37(1):54–77.

Atchadé YF, Rosenthal JS (2005) On adaptive Markov chain Monte Carlo algorithms. *Bernoulli* 11(5):815–828.

Bradlow ET, Feit EM (2018). Fusion modeling. Homburg C, Klarmann M, Vomberg AE, eds. *Handbook of Marketing Research* (Springer, Berlin).

Bronnenberg BJ, Dubé JPH, Gentzkow M (2012) The evolution of brand preferences: Evidence from consumer migration. *Amer. Econom. Rev.* 102(6):2472–2508.

Bureau of Labor Statistics (2017). Accessed September 2020, https://data.bls.gov/timeseries/LNS14000000.

Chib S, Greenberg E (1995) Understanding the Metropolis-Hastings algorithm. *Amer. Statist.* 49:327–335.

Ebbes P, Grewal R, DeSarbo WS (2010) Modeling strategic group dynamics: A hidden Markov approach. *Quant. Marketing Econom.* 8(2):241–274.

Eddy SR (1998) Profile hidden Markov models. *Bioinformatics* 14(9): 755–763.

Feit EM, Beltramo MA, Feinberg FM (2010) Reality check: Combining choice experiments with market data to estimate the importance of product attributes. *Management Sci.* 56(5):785–800.

Feit EM, Wang P, Bradlow ET, Fader PS (2013) Fusing aggregate and disaggregate data with an application to multi-platform media consumption. *J. Marketing Res.* 50(3):348–364.

Fong NM, Fang Z, Luo X (2015) Geo-conquesting: Competitive locational targeting of mobile promotions. *J. Marketing Res.* 52(5): 726–735.

Ford BM (1983) An overview of hotdeck procedures. in Madow WG, Olkin I, Rubin DB, eds. *Incomplete Data in Sample Surveys*, vol 2 (Academic Press, New York), 185–207.

Garg R, Telang R (2018) To be or not to be linked: Online social networks and job search by unemployed workforce. *Management Sci.* 64(8):3926–3941.

Gilula Z, McCulloch RE, Rossi PE (2006) A direct approach to data fusion. *J. Marketing Res.* 43(1):73–83.

Granovetter M (1973) Weak ties and strong ties. *Amer. J. Sociol.* 78: 1360–1380.

Gupta S (1991) Stochastic models of interpurchase time with time dependent covariates. *J. Marketing Res.* 28(1):1–15.

Hamilton JD (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57(2):357–384.

Hand DJ, Till RJ (2001) A Simple generalisation of the area under the ROC curve for multiple mlass classification problems. *Machine Learn.* 45:171–186.

Hauser JR, Urban GL, Liberali G, Braun M (2009) Website morphing. *Marketing Sci.* 28(2):202–224.

Hill K (2012) How Target figured out a teen girl was pregnant before her father did. *Forbes* (February 16), https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/?sh=5ae53d946668.

Hippler HJ, Schwarz N (1987) Response effects in surveys. *Social Information Processing and Survey Methodology* (Springer, New York), 102–122.

Kamakura WA, Wedel M (1997) Statistical data fusion for cross-tabulation. *J. Marketing Res.* 34(4):485–498.

Kamakura WA, Wedel M (2000) Factor analysis and missing data. *J. Marketing Res.* 37(4):490–498.

Matz SC, Netzer O (2017) Using Big Data as a window into consumers' psychology. *Current Opin. Behav. Sci.* 18:7–12.

Monaco JV, Tappert CC (2018) The partially observable hidden Markov model and its application to keystroke dynamics. *Pattern Recognition* 76:449–462.

Montgomery A, Li S, Srinivasan K, Liechty J (2004) Modeling online browsing and path analysis using clickstream data. *Marketing Sci.* 23(4):579–595.

Morwitz VG, Johnson E, Schmittlein D (1993) Does measuring intent change behavior? *J. Consumer Res.* 20(1):46–61.

Netzer O, Ebbes P, Bijmolt TH (2017) Hidden Markov models in marketing. *Advanced Methods for Modeling Markets* (Springer, Cham, Switzerland), 405–449.

Netzer O, Lattin JM, Srinivasan V (2008) A hidden Markov model of customer relationship dynamics. *Marketing Sci.* 27(2):185–204.

Qian Y, Xie H (2014) Which brand purchasers are lost to counterfeiters? An application of new data fusion approaches. *Marketing Sci.* 33(3):437–448.

Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* 77:257–286.

Romero J, Van der Lans R, Wierenga B (2013) A partially hidden Markov model of customer dynamics for CLV measurement. *J. Interactive Marketing* 27(3):185–208.

Rossi PE, McCulloch RE, Allenby GM (1996) The value of purchase history data in target marketing. *Marketing Sci.* 15(4):321–340.

Saito T, Rehmsmeier M (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 10(3):e0118432.

Scheffer T, Decomain C, Wrobel S (2001). Active hidden Markov models for information extraction. Hoffmann F, Hand DJ, Adams N, Guimaraes G, eds. *Proc. 4th Internat. Conf. Adv. Intelligent Data Anal.*, (Springer, Berlin, Heidelberg), 309–318.

Schwartz EM, Bradlow ET, Fader PS (2014) Model selection using database characteristics: Developing a classification tree for longitudinal incidence data. *Marketing Sci.* 33(2):188–205.

Schwarz N (1999) Self-reports: how the questions shape the answers. *Amer. Psych.* 54(2):93.

Schweidel DA, Bradlow ET, Fader PS (2011) Portfolio dynamics for customers of a multiservice provider. *Management Sci.* 57(3): 471–486.

Stopfer JM, Gosling SD (2013) Online social networks in the work context. Derks D, Bakker A, eds. *The Psychology of Digital Media at Work* (Psychology Press, London), 39–59.

Thompson CS, Thomson PJ, Zheng X (2007) Fitting a multisite daily rainfall model to New Zealand data. *J. Hydrology (Amsterdam)* 340:25–39.

Tourangeau R, Rips LJ, Rasinski K (2000) *The Psychology of Survey Response* (Cambridge University Press).

Trusov M, Ma L, Jamal Z (2016) Crumbs of the cookie: User profiling in customer-base analysis and behavioral targeting. *Marketing Sci.* 35(3):405–426.

Wachtel S, Otter T (2013) Successive sample selection and its relevance for management decisions. *Marketing Sci.* 32(1):170–185.

Wedel M, Kannan PK (2016) Marketing analytics for data-rich environments. *J. Marketing* 80(6):97–121.

Yamato J, Ohya J, Ishii K (1992) Recognizing human action in time-sequential images using hidden markov model. *Proc. IEEE Computer Society Conf. Comput. Vision Pattern Recognition* (IEEE, New York), 379–385.

Zarate LE, Nogueira BM, Santos TRA, Song AJM (2006) Techniques for missing value recovering in imbalanced databases: Application in a marketing database with massive missing data. *Proc. IEEE Internat. Conf. Systems Man Cybernetics*, vol. 3 (IEEE, New York), 2658–2664.

Zhang JZ, Netzer O, Ansari A (2014) Dynamic targeted pricing in B2B relationships. *Marketing Sci.* 33(3):317–337.