USING SOCIAL NETWORK ACTIVITY DATA

TO IDENTIFY AND TARGET JOB SEEKERS

Peter Ebbes*
HEC Paris

Oded Netzer
Columbia Business School

June, 2018

* Peter Ebbes is Associate Professor of Marketing, HEC Paris (email: ebbes@hec.fr). Oded Netzer is Professor of Business, Columbia Business School, Columbia University (e-mail: onetzer@gsb.columbia.edu). Peter Ebbes acknowledges research support from Investissements d'Avenir (ANR-11-IDEX-0003/LabexEcodec/ANR-11-LABX-0047) and the HEC foundation.

USING SOCIAL NETWORK ACTIVITY DATA TO IDENTIFY AND TARGET JOB SEEKERS

ABSTRACT

An important challenge for many firms is to identify the life transitions of its customers, such as job searching, being pregnant, or purchasing a home. Inferring such transitions, which are generally unobserved to the firm, can offer the firm opportunities to be more relevant to its 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 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, our modeling approach combines cross-sectional survey responses to a job seeking status question with longitudinal user activity data. Thus, in some time periods, and for some users, we observe the "true" job seeking status. We fuse the observed state information into the HMM likelihood, resulting in a partially HMM. We demonstrate that the proposed model can not only predict 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 42% increase in profits of a targeting campaign relative to the approach that was used at the time of the data collection.

1. INTRODUCTION

The increased availability of data at the customer level (Wedel and Kannan 2016) allows companies to effectively target customers based their individual characteristics (Matz and Netzer 2017), their location (Fong, Fang and Luo 2015), or their past behavior (Trusov, Ma and Jamal 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 pregnancy (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, Dubé and Gentzkow 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 wish to use its longitudinal activity data about its customer, possibly complemented by cross-sectional limited observations regarding the "true" unobserved 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 is 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

[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 firm's platform to evaluate candidates. According to the firm, a substantial part of the firm's revenue comes from targeting job seekers.

firm, as most job seekers do not publicly announce that they are seeking for a job (Garg and Telang 2017).

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, a job seeker may exhibit different forms of engagement on the social network platform such as updating her profile, more often searching for companies, or trying to grow her social network by sending invitations to connect to other users. Furthermore, a user who starts searching for a job, may exhibit increased activity on the platform compared to her own past activity. However, without knowing the job seeking state of at least a subset of the users, we cannot know to what extent the observed activity on the platform relates to job search.

To address the challenge of inferring job seeking status from users' engagement with the social network platform, we combine two sources of information: a) a large set of platform activities observed over time, such as number of visits, 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 these users at a certain point in time. In order to infer the latent state of job search, which is also transient in nature, we develop a partially hidden Markov model (PHMM) in which the latent states correspond to different levels of job seeking, and the states are partially observed through the survey responses. In our model, each state is characterized by a multivariate set of activities in 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 responded to the survey into the likelihood of a traditional HMM, making their latent states "observable" at that time. As such, the PHMM is calibrated incorporating information about job seeking status for some users at some points in

time, allowing us to make inferences regarding the job seeking states of all customers in all time periods.

We show that the proposed model can not only infer and predict 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 userbase of the social network platform, only a small subset of users can be surveyed at any time period. Hence, we demonstrate the ability of the proposed model to predict job search both for out-of-sample time periods and for out-of-sample users, who were never surveyed. We further demonstrate that targeting job seekers based on our proposed approach can lead to a 42% increase in response rates and profits relative to the approach that was used at the time of the data collection.

The contribution of our research is twofold. From a substantive point of view, we demonstrate how companies can use customers' activity data to infer the customers' latent behavior that may be 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 our context of job seeking, we uncover activities on the social network platform that are linked with job seeking, such as increased activity and strategic use of the user's social network. From a methodological point of view, we build a PHMM, which extends the traditional HMM by fusing one or more snapshots of survey data into the sequence of longitudinal activity data through the latent state component of the HMM's likelihood function. Additionally, most HMM applications in marketing leverage the latent states as a means to capture and predict the dynamics of the state-dependent behavioral outcomes (e.g., donations in Netzer, Lattin and Sriniavsan 2008, churn in Ascarza and Hardie 2013). However, this paper, like several HMM studies 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 motivates our modeling choices. Section 4 describes the main model. Section 5 presents the empirical results, and Section 6 demonstrates the use of the model for targeting purposes. In Section 7, we extend the model to generate richer managerial insights. Finally, we present the conclusions and discuss the limitations of our study in Section 8.

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 next.

2.1 Identifying 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, McCulloch and Allenby 1996; Hauser et al. 2009), commitment to or relationship with the firm (Netzer, Lattin and Srinivasan 2008; Ascarza and Hardie 2013; Romero, van der Lans and Wierenga 2013; Schwartz, Bradlow and Fader 2014; Ascarza, Netzer and Hardie 2018), price sensitivity (Zhang, Netzer and Ansari 2014), stage in the purchase funnel (Montgomery et al. 2004), attention states (Liechty, Pieters and Wedel 2003; Wedel, Pieters and Liechty 2008), learning strategies (Ansari, Montoya and Netzer 2012), portfolio of products (Schweidel, Bradlow and Fader 2011), and emotional states

(Nwe, Foo and De Silva 2003). A common theme for these papers is that they include a latent space model (often a HMM) that captures the underlying state.

HMMs are useful in situations 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 states of its users. Instead, the platform observes a host of users' activities, which may provide a noisy signal of the user's job seeking status. For example, a user who updates his or her 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 into the HMM likelihood survey responses that identify the true state for a subset of the population at a certain point in time. Netzer, Lattin and Srinivasan (2008) have 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 a way to directly fuse such survey responses into the HMM likelihood function. In that sense our work is more closely related to the limited work on PHMMs, in which some of the states are fully observed. Romero, van der Lans and Wierenga (2013) developed a PHMM to capture customer lifetime value. In their model some of the states are always observable (e.g., customer churn) and others are always unobserved (e.g., customer activity states). Similar observable churn states in HMMs can be found in Ascarza and Hardie (2013), who use "two clocks" for usage and churn,

where the churn state is observable every four time periods but the usage activity is observed in every period. Our PHMM specification and modeling approach are considerably different from the aforementioned studies because in our case, all states are unobserved, however, for some users in some time periods the specific state of the user becomes observable through his/her survey responses. 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, Decomain and Wrobel 2001), to understand precipitation and rainfall activity (Thompson, Thomson and Zheng 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 web site 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 rather 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, Ohya and Ishii 1992), speech recognition (Rabiner 1989), or DNA detection (Eddy 1998).

2.2 Identifying Job Seeking

The U.S. job search and recruiting industry in 2016 was estimated at $150 billion.[2] As for most recruiting and job search firms, an important challenge is identifying who is job searching and when.

[2] https://www.statista.com/statistics/220707/us-total-sales-in-temporary-staffing/ (last accessed, April 2018).

Using survey data, Garg and Telang (2017) provide strong empirical evidence that people are spending more 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, serve as leads or as 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 with 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 (2017) 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 websites 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 his or her 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 (2017), on the other hand, find among unemployed individuals that 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 their social network structure may be different from others. 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 that are outside their current professional network (e.g., their company).

While 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, 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 user 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.

2.3 Data Fusion

We leverage a survey conducted in a specific time period for a sample of users that identifies their job seeking status, to infer the job seeking status of a larger population of users in any given time period. In other words, we plan to fuse the information observed in the survey both cross-sectionally (to other users) and longitudinally (over time).

The idea behind data fusion is to capture the joint distribution of two (or more) observed variables for individuals for whom only a subset of the variables are observed. The fusion is based on the joint distribution of the variables for individuals from whom all variables are observed. 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, 2000) propose a statistical approach to tackle the problem of data fusion using a finite mixture approach (Kamakura and Wedel 1997) and a factor analytic approach (Kamakura and Wedel 2000). Gilula, McColluch and Rossi (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 non-parametric 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, Beltramo and Feinberg 2010), or fusing individual-level data with aggregate data (Feit et al. 2013).

The data fusion problem we face is quite different from the problems addressed in the above studies. We need to fuse survey data regarding job seeking status observed in one (or multiple) time period(s) to other time periods of the same individual *as well as* to all time periods for users that were not surveyed. Our approach for data fusion is similar in spirit to the approach taken by Kamakura and Wedel (1997) in the sense that we use a latent variable (a latent class in the case of Kamakura and Wedel and HMM latent states in our case) to fuse the observed behavior (job search status) with unobserved states. However, unlike the static nature of the latent variable in Kamakura and Wedel, our latent variable is dynamic such that we have to go beyond cross-sectional fusion and fuse information both cross-sectionally and over time.

3. DATA DESCRIPTION AND MODEL-FREE EVIDENCE

3.1 Monthly User Activity Data

We have a unique dataset from a large online social network platform that has millions of users. Our dataset 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 below). These users were members of the platform, and had at least 12 months of activity, during the data period.[3] The data contain over 60 types of user activities on the platform, such as whether the user sent or

[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.

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 her profile page, etc. 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 his/her 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 his/her company (=1) or inside his/her company (=0). Because of the long tailed nature of the continuous variables (variables 3-8 above), and to account for the possibility of a zero activity on these variables, we log-transform these variables as $f(x) = \log(1 + x)$.

Due to 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 were observable only for the first 5 month of the data period. Specifically, we observe variables 1-4 above for the entire 14 months and variables 5-9 above only for the first five months. Imbalance in data collection is quite common among firms' databases (Zarate et al. 2006). In the model section, we describe how we handle this data imbalance.

[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 aspects into a single dummy variable.

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 our data window (June 2011). We will fuse the first survey (hereafter the survey) into the model to define 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 survey responses with the social network platform activity data across users and 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:[5]

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

Following the company's classification of the response categories, we define [1]+[2] as active job seekers, [3]+[4] as passive job seekers and [5] as users who are not searching for job opportunities.

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

[5] Bolding in the response categories is for exposition purposes in the paper but not in the actual survey. This question was designed for the data provider by an external consulting firm.

actively looking for new job opportunities, 57% (=14%+43%) are passively looking for new opportunities and 21% are not looking for new opportunities.[6]

	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 (1) or Inside (0) company	Inv sent	Inv received	Conn formed	Conn invitee
[1] Actively looking	0.11	0.48	0.48	74.19	5.06	0.79	2.39	1.23	3.93	63.61
[2] Casually looking	0.10	0.25	0.36	36.57	2.33	0.88	1.32	1.32	2.81	24.38
[3] Thinking about	0.14	0.18	0.24	27.88	2.00	0.76	1.17	1.43	2.82	44.34
[4] Would discuss	0.43	0.09	0.22	25.33	1.59	0.75	1.19	1.36	2.83	41.88
[5] Not interested	0.21	0.06	0.25	21.98	1.45	0.77	1.11	1.23	2.56	29.06
Test statistic (H0: no difference b\w groups)	1009.90	227.97	65.50	26.98	24.55	3.38	10.42	1.04	5.34	2.82
P-value	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.38	0.00	0.02
N	2,814	2,814	2,814	2,814	2,814	398[1]	2,814	2,814	2,814	1,081[1]

Table 1 Comparison of the user activity during the month of the first survey across job search survey responses. Absolute numbers for activity are scaled by an unknown number.

[1] 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.

Before we investigate how one can build a predictive model of job search from the observed activity on the platform, it is useful to examine the relationship between different activities on the platform and the responses to the job seeking question in the survey.

3.3. Model-Free Evidence

The Relationship Between Job Seeking Status and Activity During the Month of the Survey

In Table 1 we compare the users' activity on the platform during the month of the survey and the users' responses to the job seeking survey question. One of the activity variables we observe is whether or not the user used the platform's job search tool. A naïve approach to

[6] At the time of our study, the U.S. unemployment rate was a little less than 10%, which closely 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).

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. We find that the job seeking status survey response significantly correlates with the use of the job search tool (chi-sq value = 227.97, P-value<0.001). Specifically, those who are actively looking for a job use the tool considerably more than other users. However, nearly 52% of those who actively search for a job according to their survey response, and nearly 75% of those who casually search for a job, did not use the job search tool during the month of the survey. Thus, while job seekers use the job search tool, many job seekers cannot be identified with this single activity. Next, we examine whether other user activities can help discriminate between active, passive and non-job seekers.

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-value = 26.98, P-value<0.001), search twice as often (F-value =24.55, P-value<0.001), and have a higher probability to update their profile page (chi-sq=65.50, P-value<0.001). We also observe that job seekers grow their social network differently from non-job 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-value=5.34, P-value<0.001). In addition, we find that job seekers were more likely to send invitations to connect, trying to expand their network (F-value= 10.42; P-value<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-value=1.04, P-value = 0.38). Thus, there is an asymmetry between invitations sent and invitations received across the various job seeking categories.

Lastly, one could ask whether users strategically expand their network for job search purposes. To investigate this, we examine whether active job seekers, relative to passive and non-job seekers, were more likely to connect to users who are well connected. We find that job seekers seem to be strategic in growing their network, connecting to other users that have relatively more connections than the users to whom passive and non-job seekers are connecting to (F-value=2.82, P-value=0.02).

Longitudinal Analysis of Relationship between Job Seeking Status and Activity

The analysis described above 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 one single activity cannot accurately reveal the user's job seeking status. Hence, a multivariate approach to characterize job seeking behavior may be 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 prior to 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 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 is 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.

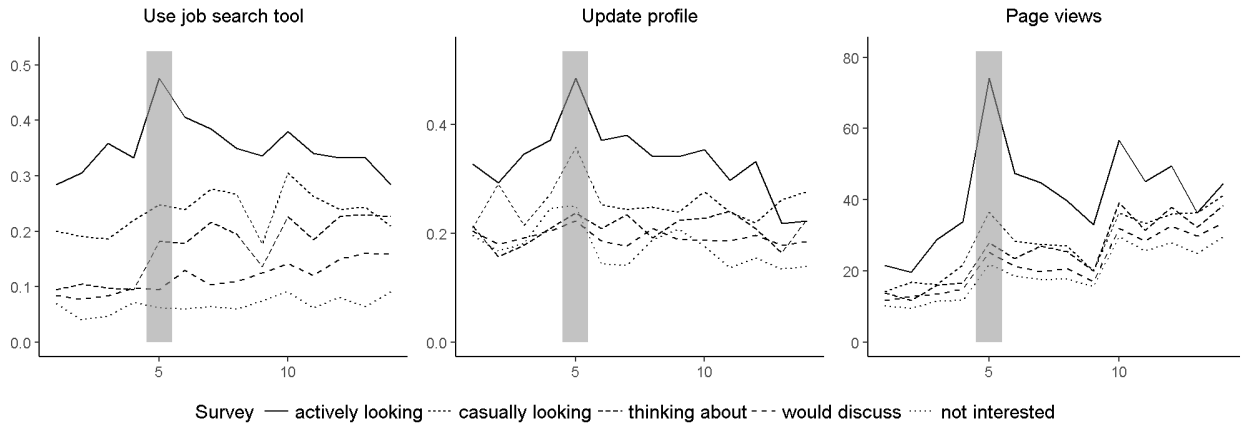

Figure 1. Average monthly activity levels of probability to use the job search tool, to update the profile, and the number of page views during the observation period. The survey was fielded in month 5 (shaded area). Absolute numbers for activity are scaled by an unknown number.

Several observations regarding Figure 1 are noteworthy. First, we observe that activity on the platform is increasing over time. Particularly, the average number of page views and the use of the job search tool increase over time. To account for such an increase, and to distinguish it from job search patterns, we include the number of unique visitors to platform[7] 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 an active or casual job seeker but not for other users who report to be not job seeking in month 5. The increase in profile update activity seems to start prior to month 5, as some of

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

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 a different behavior on the platform than non-seekers, and that behavior should be characterized by a multivariate set of activities. Second, the activity of job seekers changes over time, presumably when their (latent) job seeking status changes. Thus, the users' activity level and its 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, we need to fuse in our model 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, Ebbes and Bijmolt 2017). As argued above, this class of models suits our research problem and data well. We observe users' activities on the platform, which serve as noisy signals of the latent variable of interest – the users' job seeking states. However, it is important to model the dynamics in the job seeking state, because users transition in and out of job seeking over time.

4.1 A Three-state PHMM of Job Seeking with Data Fusion for the Survey Responses

We initially consider three states of job search, following the company's categorization of types of job seeker: non-job seeker, passive job seeker, and active job seeker, and discuss

extensions to more than three states in Section 7. Hence, we consider a HMM with three latent states of job search, say, S_{it}, with a finite state space $\{1,2,3\}$, for user $i = 1,2,\dots,N$ in month $t = 1,2,\dots,T$. Each user can be in one of the three states in a given month, and transition among states over time. What we observe is multivariate user activity data, Y_{it}, where Y_{it} is a $P \times 1$ vector of P user activities (e.g., profile update, total number of searches etc.). In a HMM, we assume that the probability distribution of Y_{it} depends on S_{it}. For example, a user in the active job seeking state may be more likely to use the job search tool or view more pages relative to a user in a passive or non-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 reveals the unobserved state S_{it} at the period of the survey and we can use this information to update the likelihood function corresponding to the exact path 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 in calibrating 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. In the extreme, if the company had surveyed all users in every time period, then we would have a standard Markov model (e.g., Leeflang et al. 2015). Of course, collecting such data is largely impractical. Figure 2 schematically illustrates the PHMM in our application.

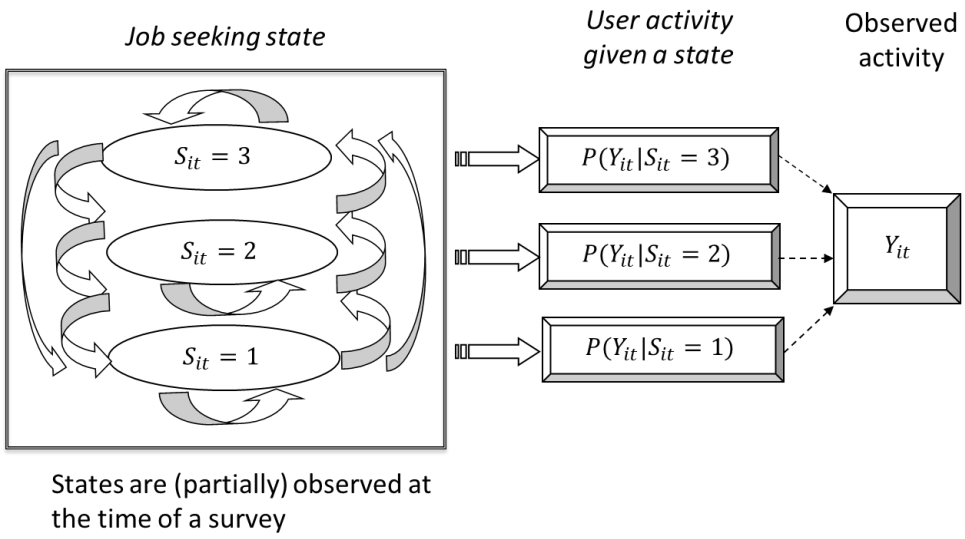

Figure 2 Schematic representation of PHMM for job search and user activity.

Formally, the model consists of three main components: the initial state distribution, the transition probabilities and the state-dependent activity distributions. The initial state distribution specifies the job seeking state at the beginning of the data period. This distribution is a discrete distribution, given by $\pi = \{\pi_1, \pi_2, \pi_3\}$ where $\pi_j = P(S_{i1} = j)$, for $j = 1,2,3$, which we estimate through a vector of 2 parameters (the probabilities sum to 1). The transition probabilities describe the stochastic process S_{it}. As is common for HMMs, this process is assumed to satisfy the Markov property so that the user's job seeking state in month t, only depends on the user's job seeking state in month $t-1$ and does not depend on the months before $t-1$, i.e., $q_{kj} = P(S_{it} = j|S_{it-1} = k)$, for $j, k = 1,2,3$. We represent these probabilities by a 3×3 transition probability matrix, Q. Lastly, the state-dependent activity distributions in a HMM describe the observed activities, given the user's state S_{it}, i.e. $m_{itj} = P(Y_{it}|S_{it} = j)$ for $j = 1,2,3$. We observe several types of activity, specifically, P_1 discrete activities (e.g., the user updated her profile page) and P_2 continuous activities (e.g., number of page views). Hence, Y_{it} is a $P \times 1$ vector, with $P =$

$P_1 + P_2$. As mentioned above, some types of activity are only observable for the first 5 months of the data period, which we accommodate by varying the length of the vector Y_{it}.

We model the discrete activities as a binary logit model and the continuous activities as a Tobit-regression model (the continuous activities are bounded at 0). The coefficients of these models are state-dependent. We write the state dependent probabilities as a 3x3 diagonal matrix, M_{it}, with the diagonal elements representing the conditional probabilities $m_{itj} = P(Y_{it}|S_{it} = j)$, with $j = 1,2,3$. 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 (π, Q, and M).

We first discuss the general form of the HMM likelihood function, ignoring the fact that for some users in some time periods we observe their "true" job search state. We will then discuss how this information can be fused into the HMM, resulting in a PHMM. The probability of observed data for user i, given the user-specific vector of random intercepts α_i and the vector of fixed-effect parameters θ, is given by:

$$P(Y_{i1}, Y_{i2}, \dots, Y_{iT}|\alpha_i, \theta) = \pi_i M_{i1} Q_i M_{i2} Q_i \dots Q_i M_{iT} \iota, \quad (1)$$

where the vector α_i contains the user specific random intercepts for π, Q, and M, and ι is a 3×1 vector of ones. Specifically, $\alpha_i = (\alpha_i^{\pi}, \alpha_i^{Q}, \alpha_i^{M})$, where $\alpha_i^{\pi} = (\alpha_{i1}^{\pi}, \alpha_{i2}^{\pi})'$ is a 2×1 vector, $\alpha_i^{Q} = vec(\mathrm{A}_i^{Q})$, A_i^{Q} a 3×2 matrix with as (k, j)-th element α_{ikj}^{Q}, and α_i^{M} a vector of random intercepts for the continuous activity variables.[8] We assume a multivariate normal distribution for the

[8] 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).

upper-level model of the random intercepts, i.e. $\alpha_i \sim N(0, \Sigma_\alpha)$. The elements of the initial state distribution are:

$$\pi_{ij} = P(S_{i1} = j|\alpha_i^\pi, \theta) = \frac{\exp\,(\tau_j + \alpha_{ij}^\pi)}{1 + \exp(\tau_1 + \alpha_{i1}^\pi) + \exp\,(\tau_2 + \alpha_{i2}^\pi)}, \quad (2)$$

for $j = 1,2$ and $\pi_{i3} = P(S_{i1} = 3) = 1 - \pi_{i1} - \pi_{i2}$. The parameters τ_j are the baseline logit threshold parameters, and the sum $\tau_j + \alpha_{ij}^\pi$ is the threshold value for individual i to be in state j, for $j = 1,2$. The elements of the transition probability matrix Q_i are:

$$q_{ikj} = P\left(S_{it} = j \middle| S_{it-1} = k, A_i^Q, \theta\right) = \frac{\exp\,(\phi_{kj} + a_{ikj}^Q)}{1 + \exp(\phi_{k1} + a_{ik1}^Q) + \exp\,(\phi_{k2} + a_{ik2}^Q)}, \quad (3)$$

for $j = 1,2$ and $q_{ik3} = P(S_{it} = 3|S_{it-1} = k) = 1 - q_{ik1} - q_{ik2}$, and $k = 1,2,3$. The thresholds ϕ_{kj} are the baseline intercepts for the logit probability that a user is transitioning from state k to state j in a given time period, for $j = 1,2$, and $k = 1,2,3$.

The state-dependent probability matrix M_{it} for the user activity is a diagonal matrix containing the following elements:

$$m_{itj} = P\left(Y_{it} \middle| S_{it} = j, \alpha_i^M, \theta\right) = \left(\prod\nolimits_{p=1}^{P_1} P(Y_{itp}|S_{it} = j, \theta)\right) \times \left(\prod\nolimits_{p=P_1+1}^{P_1+P_2} f(Y_{itp}|\, S_{it} = j, \alpha_i^M, \theta)\right), \quad (4)$$

for $j = 1,2,3$. The probability model for the P_1 discrete variables is:

$$P\left(Y_{itp} = 1 \middle| S_{it} = k, \theta\right) = \frac{\exp\,(\delta_{0pk} + \delta_{1p} Z_t)}{1 + \exp\,(\delta_{0pk} + \delta_{1p} Z_t)}, \quad (5)$$

with δ_{0pk} being the logit intercept for observing activity p in state k, $p = 1,2,\ldots,P_1$, $k = 1,2,3$, and δ_{1p} is the regression coefficient for observed activity p for the control variable Z_t, which is the unique number of visitors to the platform to capture general aggregate trends in activity during the data period.

The continuous variables that are observed for the whole data period are modeled as a Tobit regression, including a user-specific random intercept to capture base-line activity, and the unique number of visitors as control variable. For the p-th continuous variable we have

$$f\left(Y_{itp}\middle|S_{it}=k,\alpha_i^M,\theta\right)=Tobit(\mu_{itpk},\sigma_{pk}^2), \tag{6}$$

with

$$\mu_{itpk}=\beta_{0pk}+\beta_{1p}Z_t+\alpha_{ip}^M, \tag{7}$$

where β_{0pk} is the intercept of the p-th variable in state k, $p=1,2,\dots,P_2$, β_{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 Tobit model for activity variable p and state k. As mentioned above, we log transform the monthly activity levels.

The model in Equations (1)—(7) represents a standard HMM. 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. Similarly, only transitions out of state s are allowed into any state between period t and period $t+1$. This will constrain the transition probability matrices for this user going into and out of time period t. We define $Q_{i,\cdot\to s}^t$ as a 3×3 matrix of zeros where the s-th column is the s-th column of Q_i and $Q_{i,s\to\cdot}^t$ as a 3×3 matrix of zeros where the s-th row is the s-th row of Q_i. For example, suppose the user indicates she is in the active job seeking state (State 3) at time period t. Now we can

constrain the transition going into state $s = 3$ in period t (left matrix in Equation (8)) and the transition going out of state $s = 3$ in period t (right matrix in Equation (8)).

$$Q^t_{i,\cdot\to s=3} = \begin{bmatrix} 0 & 0 & q_{i13} \\ 0 & 0 & q_{i23} \\ 0 & 0 & q_{i33} \end{bmatrix} \text{ and } Q^t_{i,s=3\to\cdot} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ q_{i31} & q_{i32} & q_{i33} \end{bmatrix} \quad (8)$$

We then modify the likelihood function in Equation (1) to include the observability of the latent state for user i when she responds to the survey in time period t as:

$$P(Y_{i1}, Y_{i2}, \dots, Y_{iT} | \alpha_i, \theta) = \pi_i M_{i1} Q_i M_{i2} Q_i \dots Q_i M_{it-1} Q^t_{i,\cdot\to s} M_{it} Q^t_{i,s\to\cdot} M_{it+1} Q_i \dots Q_i M_{iT} \iota. \quad (9)$$

The model in Equation (9) could be easily modified if the researcher observes the true state in multiple time periods. The likelihood function in (9) is a type of a PHMM, in which the researcher observes the latent state in some but not all time periods. Furthermore, the 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 (e.g., 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 useful for applications in which state recovery, as opposed to outcome predictions, is the main objective of the modeling effort.

4.2 Model Estimation Approach

We use a Bayesian framework to estimate our PHMM and incorporate cross-user heterogeneity (e.g., Ebbes, Grewal and DeSarbo 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). We note that fusing the observed survey responses into the HMM likelihood

greatly helps in keeping the labels sorted over the course of the MCMC sampling. We did not find any label switching in our model estimates. We present in Web Appendix A the details of the MCMC algorithm used.

5. EMPIRICAL APPLICATION

We calibrate the PHMM 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 first survey (month 5 of the data window) into the PHMM 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 job seeking for out-of-sample users, we split the data into a calibration sample ($N_c = 400$) and a validation sample ($N_v = 91$).

5.1 PHMM Posterior Estimates

We estimated the proposed PHMM using the Bayesian MCMC approach described in Section 4.2. Table 2 reports the posterior mean and posterior standard deviation of the parameters of the three components of the PHMM (π, Q, and M). For ease of interpretation we transformed the working parameters (α_i and θ) into posterior probabilities for the discrete variable in M, the initial state and the transition matrix, and the anti-log of the expected values for the continuous variables in M. The trend parameters are reported at the working parameter level.[9]

[9] The posterior mean and standard deviation of the working parameters is available from the authors upon request.

	State			
	Non Seeking	Passive	Active	Trend parameter
Profile updates (dum)	0.04 (0.01)	0.17 (0.01)	0.56 (0.01)	-0.008 (0.003)
Job searched (dum)	0.01 (0.00)	0.10 (0.01)	0.65 (0.03)	0.011 (0.004)
Total searches	5.50 (1.18)	3.94 (0.21)	18.15 (1.04)	0.014 (0.002)
Pageviews	12.63 (0.80)	68.46 (2.10)	217.24 (6.44)	0.017 (0.001)
More invitations outside company (dum)	0.76 (0.20)	0.80 (0.08)	0.92 (0.04)	0.045 (0.048)
Invitations sent	3.81 (0.89)	3.05 (0.25)	7.33 (0.48)	0.001 (0.012)
Invitations received	2.04 (0.21)	2.15 (0.08)	3.00 (0.17)	-0.002 (0.007)
Connections formed	2.84 (0.37)	3.17 (0.13)	7.91 (0.41)	-0.003 (0.007)
Number of connections of invitee	160.64 (218.14)	58.29 (11.94)	126.45 (18.06)	0.054 (0.018)
Initial state distribution	0.42 (0.03)	0.36 (0.03)	0.22 (0.02)	
Transition matrix				
From non-seeking to…	0.48 (0.02)	0.36 (0.03)	0.15 (0.02)	
From passive to…	0.19 (0.01)	0.62 (0.02)	0.20 (0.02)	
From active to…	0.16 (0.02)	0.41 (0.03)	0.43 (0.03)	

Table 2. Posterior means and standard deviations (in parentheses).

There are several important observations to note from the posterior estimation results in Table 2. First, we see that the estimates are consistent with the model-free evidence (Section 3.3). That is, job seekers 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 are actively searching 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 7.33/3.00 = 2.44) compared to non-seekers and passive seekers, for whom this ratio is more balanced. Consequently, the active job seekers tend to form more connections, generally connections who are well connected themselves, suggesting that there is some strategic behavior among job seekers in the way they grow their network. The transition matrix demonstrates that the passive state is most sticky, followed by the non-seeking state. If a user is in the active job seeking state in month t, then the probability that in the next time period (s)he is again in the active job seeking state is 0.43. The stickiness of the active job seeking state implies a duration of about 1.7-1.8 months of

active job seeking. This result is fairly consistent with the reported median duration of unemployment of approximately 10 weeks.[10] See Web Appendix B for a discussion of the posterior results of the heterogeneity distribution.

5.2 Posterior Predictions of Job Search

To identify job seekers, the company needs to predict the job seeking status of the entire user base over time, as it is impossible to survey all users at every time period. Thus, the company needs to predict the job seeking status of users who never responded to a job seeking survey as well as the status of users who responded to a survey in different time periods. To test the model for such prediction scenarios, we consider predicting the survey response of out-of-sample users – users ($N_v = 91$), who were not used for model calibration, and predicting users in out-of-sample time periods – predicting the responses to the validation survey, which occurred one month after the end of the calibration data window. Table 3 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 behaviors (M) in future periods, but rather to predict the latent states of the users.

		Time	
		Month 5 – Survey 1	Month 14 – Survey 2
Cross section	Calibration sample ($N_c = 400$)	**In-sample users & in-time period** No predictions are made as the first survey is deterministically fused into the PHMM for the calibration sample.	**[1] In sample users, out-of-time period** Predict job seeking status in month 14 for users whose responses to Survey 1 were used to calibrate the model.
	Holdout sample ($N_v = 91$)	**[2] Out-of-sample users & in-time period** Predict job seeking status in month 5 for a hold-out sample of users at the time period of the first 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.

Table 3. Schematic overview of the prediction analyses

[10] https://www.bls.gov/opub/ted/2011/ted_20110602.htm (last accessed: April 2018).

Thus, we consider three types of holdout predictions (Table 3):

(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 (other) users were surveyed. We use only the observed activity during the first five 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 (which in our case contains millions of users). Hence, this scenario is the "cleanest" and most practical prediction scenario to test our model.

We note that, by definition, the model fit is perfect for the calibration sample in month 5 when the survey was run, as the user responses to the survey were deterministically fused into the PHMM.

We use the model's state predictions and the job seeking status reported in the surveys to calculate predictions. The predictions in month 5 are validated with responses to the first survey; the predictions in month 14 are validated with responses to the second survey. In order to compute the posterior probabilities of state membership for each calibration user in month 14, i.e. $P(S_{i14}|\alpha_i, \theta, Y_{i1}, Y_{i2}, \dots, Y_{i14})$, $i = 1,2, \dots, N_c$, we use the filtering approach (Netzer, Ebbes and

Bijmolt 2017) in each step of the MCMC sampler. We use the "max probability rule" on the posterior means to assign each user to a job seeking state.

A challenge arises in computing posterior state membership probabilities for the holdout sample users ($N_v = 91$), because we do not have estimates for the individual-level parameters (α_i). We therefore use the following procedure. Taking $\theta = \bar{\theta}$ fixed at the posterior mean estimated from the calibration sample, we run the observed activity in the first 5 months of the data of each validation user ($N_v = 91$) through the MCMC sampler, to generate a posterior sample of size L of random intercepts α_i^l, $i = 1,2, \dots, N_v$, $l = 1,2, \dots, L$. Next, using the same filtering approach, we calculate $P(S_{i5}|\alpha_i^l,\ \bar{\theta},\ Y_{i1}, Y_{i2} \dots, Y_{i5})$ and $P(S_{i14}|\alpha_i^l,\ \bar{\theta},\ Y_{i1}, Y_{i2} \dots, Y_{i14})$, for each $l = 1,2, \dots, L$. After computing the posterior means across the L draws, we use the "max probability rule" to assign each holdout user to a job seeking state.

We compare the predictions of the PHMM to an observed state benchmark model: an ordered logit model with three categories (non, passive and active job seeking) calibrated on the survey responses in month 5, using as covariates the same (nine) variables that were used to calibrate the PHMM. We use the observed current and lagged activities of the users in months 4 and 5 to predict the job seeking status in month 5, and the current and lagged observed activity in months 13 and 14 to predict the job seeking status in month 14. Similar to the PHMM, the observed state ordered logit benchmark model includes dynamics via the lagged observed activities as covariates. Thus, the ordered logit model is a strong contender as it fits directly to the survey responses as a function of current and past activity.[11]

We compute three metrics (Jaccard index (J), the Fowlkes–Mallows index (FM), and the Classification success index (CSI)) to evaluate the job seeking status predictions of the proposed

[11] Estimates of the ordered logit model are provided in the Web Appendix C.

PHMM and the ordered logit model. Because our interest is in predicting active job seeking (as opposed to non-job seeking or passive job seeking), we chose metrics which employ a loss function that focusses on the prediction of active job seeking. In order to calculate these metrics, we distinguish between active job seeking and the combination of passive and non-job seeking.[12] The prediction results for the three metrics are given in Table 4.

			Time			
			Month 5 – Proposed	Survey 1 Ord. logit	Month 14 – Proposed	Survey 2 Ord. logit
Cross section	Calibration sample ($N_c = 400$)	J			0.21	0.15
		FM			0.35	0.28
		CSI			-0.30	-0.40
	Holdout sample ($N_h = 91$)	J	0.28	0.26	0.23	0.15
		FM	0.45	0.44	0.37	0.29
		CSI	-0.08	-0.08	-0.26	-0.33

Table 4. Results holdout predictions for the proposed PHMM and the ordered logit model. Performance metrics (J, FM, and CSI) indicate model performance to predict whether a user is an active job seeker in month 5 and month 14. Higher numbers indicate better performance.

First, we observe from Table 4 that predictions of the job seeking status (Survey 1) of the hold-out sample in month 5 are best and fairly similar for the two models. The relatively good predictions of the ordered logit model in month 5 can be expected as the logit model fits directly to that month's survey responses for the calibration sample. However, the prediction results of month 14 show an important disadvantage of the logit model, when the aim is to predict the job seeking status in a future period (month 14), the proposed PHMM out predicts the ordered logit model, possibly due to the PHMM's ability to capture dynamics in a more flexible way. Thus, in order to improve the performance of the ordered logit model, one would need to survey more often the userbase. Interestingly, according to the J and FM metrics the predictive ability of the

[12] See Web Appendix D for details of the calculation of the three metrics.

PHMM for the holdout period month 14 is fairly comparable to its predictive ability for the calibration period month 5. Thus, unlike the ordered logit model, the PHMM is able to predict users' job seeking status both out of sample and out of time.

Thus, the proposed PHMM outperforms the ordered logit benchmark model in predicting active job seekers. We next explore how well the proposed PHMM performs in predicting job seeking duration, and how the platform could leverage these predictions to target active job seekers.

5.3 Predicting the Duration of Job Search

Thus far we have focused on predicting whether or not a user is an active job seeker in a particular month. However, by the nature of the latent states, and the transitions among them, the PHMM should also be able to predict how long a user has been searching and when the user transitioned into the job seeking state. In order to test the ability of the proposed model to capture job search duration, we also asked respondents in the validation survey, that was fielded shortly after month 14, *how long* they had been job seeking. We use these survey responses as additional validation for the proposed model. We emphasize that the validation survey was not used for calibrating the PHMM.

We use the proposed PHMM to predict the job seeking state of the user in months 8 through 14. We then split the users by their survey response to the validation survey into two groups: those who were actively searching and those who were not searching for a job. Those that indicated they were actively job searching, were further split into two groups of job search duration: 1) those that were actively searching for at most three months, and 2) those that were actively searching for more than three months.

If the PHMM predicts job searching well, we should expect to 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 job seeking state in month 14 relative to users who are not seeking for a job. Moreover, we should expect to see that users who indicate in the validation survey that they have been searching for a job for up to three months should transition from a low probability of being in the job seeking state up to month 11 to a higher probability of being in the job seeking state after month 11. The state prediction of the PHMM and the ordered logit model (for comparison) are provided in Figure 3.

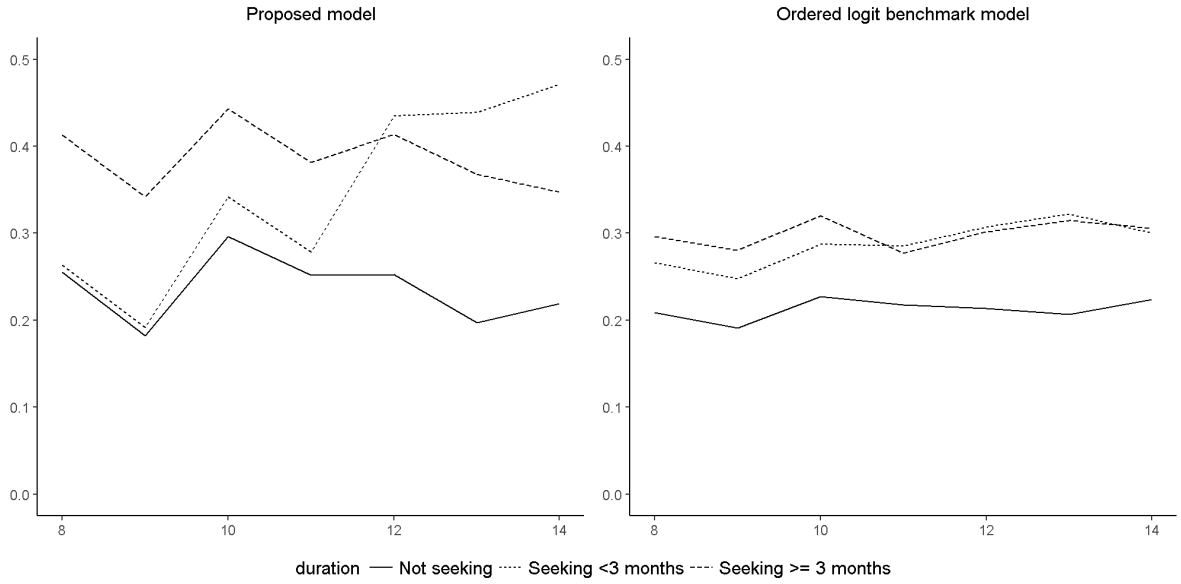

Figure 3. Average probabilities of being in the active job seeking state for the months 8—14 for the PHMM (left) and ordered logit model (right). Dashed line: the average probability of being in the active job seeking state for users that indicated in the validation survey that they were actively searching 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.

Several interesting insights can be obtained from Figure 3. The dashed line indicates the average probability of being in the active job seeking state for users who stated in the validation survey that they were actively searching for more than 3 months and the dotted line indicates the

average probability of being in the active job seeking state for users who stated in the validation survey that they were actively searching for up to 3 months. The solid line represents the average probability of being in the active job seeking state for users who indicated they were not searching for jobs in the validation survey. Consistent with the results in Table 3, the PHMM clearly separates job seekers from non-job seekers in month 14. 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) in the validation survey. Note that the separation in month 14 is less clear for the ordered logit model (right figure), 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 very well in, not only predicting who is job seeking, but also in predicting *when* the user transitioned to a job seeking state. Specifically, for those users who indicate that they were job seeking for up to three months (dotted line), the PHMM shows a transition from a behavior similar to non-job seekers prior to 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 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, relatively to those who state they were not job seeking in the validation survey (solid line). Unlike the PHMM, the ordered logit model is not able to pick up this signal, as the dotted and dashed lines are similar throughout the six months.

The results in Figure 3 demonstrate an important benefit of the proposed PHMM – it can detect changes over time in users' likelihood of being in a job seeking state, and can therefore be

used to early detect changes in the user's job seeking status. 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 model as job seekers would lead to higher response rates and higher profits.

The data used to calibrate and validate the model (described in Section 3), did not include exposure and responses to d-mails. However, we obtained from the social network platform a second user activity dataset that includes, in addition to the user activity on the platform, information on whether and when the user received a d-mail and whether s/he responded to it. This dataset also allows us to obtain convergent validity for our model and the ability of the model to capture job seeking status. The second activity dataset includes 1,621 users from whom

we observe their activity on the platform during the 12 month time period 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, which is fused into the PHMM.[13]

We use the same model and estimation procedure described in Section 3 to estimate the proposed PHMM on this second dataset. The interpretation of the states and model estimates are consistent with those found for the first dataset in Section 5 (See Web Appendix E for the posterior estimates of the parameters of the PHMM for the second dataset).

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, 864 d-mails were sent during the data period with an average of 0.53 d-mails per user across the 12 months.

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 5). It can be seen 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 job seekers. One possible explanation is that senders have no obvious way of recognizing who is an active job seeker on the platform. These preliminary analyses suggest that 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 that are not responded to.

[13] For this sample, we observe a slightly different set of activities compared to the first dataset. Specifically, we observe whether the user viewed any jobs on the platform, whether the user updated her 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.

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.

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	5.0%	32.5%
Active	4.6%	33.3%
N (Sample size)	1,621	72

Table 5. D-mails received and responded to in the month of the job seeking survey based on the users' responses to the survey.

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 them based on:

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 state according to the proposed model, and subsequently select the 100 user 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 state, the response rate increases to 52%, resulting in a profit for the platform of $5.2 per d-mail sent. This corresponds to a lift in profit of 42%. Given the number of d-mails sent on the platform every month, such a lift in profit could have substantial financial implications.

7. A PHMM WHERE THE NUMBER OF STATES DOES NOT EQUAL THE NUMBER OF SURVEY CATEGORIES

Thus far we proposed a PHMM to capture the dynamics of activity on the platform and link it to job seeking by fusing the survey information into the model likelihood. However, it may be restrictive to believe that each type of job seeker is represented by only one latent state of activity. It is entirely possible that, for example, an active job seeker exhibits different types of activity (e.g., those who use the platform as a window dressing to showcase themselves for offline job search versus those who actually use the platform to find jobs). In this section we expand the proposed PHMM to allow for multiple activity states to correspond to a particular job seeking survey response category.

In order to do so, we use the following notation. For the three state PHMM above, we indicate the fusion of the three survey response categories with the first, second and third latent states (up to label switch) as {N,P,A}, where N stands for non-job seeking, P for passive job seeking, and A for active job seeking. Extending this notation to a four-state PHMM, we can define the following three options to fuse the three categorical survey responses and the four latent states of the PHMM: {N,N,P,A}, {N,P,P,A}, {N,P,A,A}.[14] Similarly, the five state PHMM can have six options. The question that arises is how to fuse the survey responses into the PHMM likelihood, when the same survey response could correspond to more than one state.

We need to make the following modifications to the proposed three state PHMM to fuse the survey responses into PHMMs with more than three states. The vector of initial state probabilities π_i now becomes a $1 \times K$ vector, where K is the number of states, and the matrices

[14] We assume that we observe at least one response to each of the three job seeking categories N, P, and A. Therefore, we restrict our PHMM options to include at least one of each N, P, and A. If this would not be the case, then one can remove some of the job seeking response categories.

Q_i and M_i become $K \times K$ matrices. The elements of these vectors and matrices are defined as before and can be extended straightforwardly to the more general case of K states. We extend the 3×3 constraint transition probability matrix in Equation (8) to a general $K \times K$ constraint matrix, where multiple states now correspond to the same survey job status category. For example, for the case of a four state PHMM with two states for active job seeking (e.g., NPAA), and a user who responded being an active job seeker in the survey in month t, Equation (8) would be replaced by:

$$Q^t_{i,.\to s=3,4} = \begin{bmatrix} 0 & 0 & q_{i13} & q_{i14} \\ 0 & 0 & q_{i23} & q_{i24} \\ 0 & 0 & q_{i33} & q_{i34} \\ 0 & 0 & q_{i43} & q_{i44} \end{bmatrix} \text{ and } Q^t_{i,s=3,4\to .} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ q_{i31} & q_{i32} & q_{i33} & q_{i34} \\ q_{i41} & q_{i41} & q_{i43} & q_{i44} \end{bmatrix}. \quad (10)$$

The likelihood function in Equation (9) remains as before with no added complexity. We can use model selection criteria to choose both the number of states and the mapping from the observed survey response categories to the PHMM latent states. We provide further details of the model selection procedure and holdout comparisons of PHMMs with different number of states and mapping to the job seeking responses in Web Appendix F. While the five state NPAAA model fits the data best, the prediction ability of the three best fitting PHMMs (NPA, NPAA, NPAAA) are fairly similar. Tables 6 and 7, present the posterior estimation results for the NPAA and NPAAA models, which were identified by the model selection criteria as the best models to map the observed survey responses into multiple PHMM states. We use the same dataset and variables as in section 5.1 for model estimation.

State	Non Seeking	Passive	Active 1	Active 2	Trend parameter
Profile updates (dum)	0.04 (0.01)	0.14 (0.01)	0.35 (0.03)	0.66 (0.03)	-0.010 (0.003)
Job searched (dum)	0.01 (0.00)	0.07 (0.01)	0.39 (0.03)	0.75 (0.04)	0.007 (0.004)
Total searches	5.68 (1.39)	3.45 (0.21)	8.34 (0.53)	31.35 (2.45)	0.009 (0.002)
Pageviews	11.17 (0.84)	53.45 (2.22)	116.19 (3.85)	334.26 (13.93)	0.014 (0.001)
More invitations outside company (dum)	0.76 (0.20)	0.80 (0.10)	0.91 (0.10)	0.91 (0.04)	0.046 (0.050)
Invitations sent	3.67 (0.89)	3.76 (0.41)	2.49 (0.17)	17.07 (1.29)	0.009 (0.011)
Invitations received	2.08 (0.24)	2.10 (0.10)	2.56 (0.15)	3.42 (0.25)	0.002 (0.007)
Connections formed	2.95 (0.41)	3.12 (0.14)	3.90 (0.21)	16.64 (0.98)	0.003 (0.007)
Number of connections of invitee	131.44 (96.89)	64.18 (12.65)	51.45 (11.37)	453.89 (49.34)	0.045 (0.016)
Initial state distribution	0.40 (0.03)	0.27 (0.03)	0.25 (0.03)	0.08 (0.01)	
Transition matrix					
	0.44 (0.03)	0.31 (0.03)	0.16 (0.02)	0.08 (0.02)	
	0.15 (0.01)	0.52 (0.02)	0.26 (0.02)	0.07 (0.01)	
	0.19 (0.03)	0.37 (0.03)	0.35 (0.02)	0.09 (0.01)	
	0.11 (0.04)	0.27 (0.05)	0.25 (0.04)	0.37 (0.03)	

Table 6. Posterior means and standard deviations (in parentheses) for the four state PHMM (NPAA).

State	Non Seeking	Passive	Active 1	Active 2	Active 3	Trend parameter
Profile updates (dum)	0.06 (0.01)	0.32 (0.03)	0.01 (0.00)	0.31 (0.03)	0.66 (0.03)	-0.011 (0.003)
Job searched (dum)	0.03 (0.01)	0.14 (0.01)	0.00 (0.00)	0.42 (0.03)	0.76 (0.03)	0.003 (0.004)
Total searches	2.64 (0.25)	5.39 (0.45)	0.11 (0.25)	7.78 (0.55)	29.66 (2.14)	0.008 (0.002)
Pageviews	39.83 (1.98)	74.23 (4.26)	7.40 (0.63)	118.98 (3.80)	325.99 (11.97)	0.011 (0.001)
More invitations outside company (dum)	0.73 (0.17)	0.89 (0.08)	0.15 (0.37)	0.88 (0.08)	0.91 (0.04)	0.012 (0.046)
Invitations sent	2.93 (0.48)	3.91 (0.39)	0.06 (0.07)	2.83 (0.20)	18.33 (1.44)	0.008 (0.011)
Invitations received	2.01 (0.10)	2.25 (0.18)	1.70 (0.24)	2.97 (0.17)	3.47 (0.27)	0.002 (0.007)
Connections formed	2.70 (0.14)	3.42 (0.23)	1.52 (0.27)	5.43 (0.25)	17.47 (1.13)	-0.001 (0.006)
Number of connections of invitee	65.43 (26.27)	62.70 (14.53)	1,764.16 (1,543.89)	62.44 (14.65)	478.53 (53.91)	0.049 (0.015)
Initial state distribution	0.18 (0.03)	0.21 (0.03)	0.36 (0.02)	0.18 (0.02)	0.07 (0.01)	
Transition matrix						
	0.36 (0.03)	0.26 (0.03)	0.13 (0.02)	0.21 (0.03)	0.04 (0.01)	
	0.27 (0.03)	0.28 (0.04)	0.18 (0.02)	0.18 (0.03)	0.09 (0.02)	
	0.18 (0.02)	0.29 (0.03)	0.40 (0.03)	0.07 (0.02)	0.06 (0.01)	
	0.30 (0.05)	0.20 (0.04)	0.01 (0.00)	0.37 (0.04)	0.13 (0.02)	
	0.20 (0.04)	0.11 (0.03)	0.04 (0.02)	0.27 (0.04)	0.38 (0.03)	

Table 7. Posterior means and standard deviations (in parentheses) for the five state PHMM (NPAAA).

We observe an interesting pattern when we increase the number of states from three to four by adding an additional job seeking state (the NPAA model). While the posterior results for the non-seeking and passive states are similar to the those of the model reported in Section 5.1, two types of active job seekers emerge. First, users in the Active 1 state are type of job seekers that use the platform to search for jobs, and exhibits fairly high activity levels for searches and page views. However, the social activity of these users is not different from that of the non- and passive seekers. In contrast, users in the Active 2 state, are not only more active than all the other types of users, but are in particularly leveraging the social network aspect of the platform, possibly to seek for a job. That is, they send many more invitations to connect (but only receive slightly more, on average, than users in the other states), they grow their network faster, and they attempt to connect to users that have many connections. Examining the transition probabilities, we find that users in the passive job seeking state are much more likely to move to the Active 1 state (26%) than to the Active 2 state (7%).

Adding a third active job seeking state to have a total of five states (the NPAAA model), we find that the Active 2 state in the NPAAA model is similar to the Active 1 state in the NPAA model, and the Active 3 state in the NPAAA model is similar to the Active 2 state in the NPAA model. However, a third active job seeking state emerges (Active 1). The model captures a job seeking state with very low average activity levels that are similar and are often even lower than the activity level of non-job seekers. These job seekers are probably not using the platform to job search, potentially searching for jobs via other means. Examining the transition probability matrix, it appears that users who are job seeking but are not using the platform to search for a job (users in the Active 1 state) are not likely to start adopting the platform for their job search (low transitions between the Active 1 state and the Active 2 and 3 states). Similarly, job seekers who

are using the platform to search for a job (users in the Active 2 and 3 states) are not likely to stop using it for job search (low transitions between the Active 2 and 3 states and the Active 1 state).

Thus, we find that allowing for multiple activity states to correspond to a single job seeking response category, can help to identify different types of job seekers with respect to how they use the social network platform to search for a job. Such insights can be used by the social network platform to better target different features of the platform to different users.

8. 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 the observed activities themselves but rather by consumers' latent states and/or traits, such as job seeking, expecting a child, relocation, etc. In order to successfully target customers, it is important to identify the customer latent state from their observed behavior. The targeting of customers may be particularly important during periods of transition from one state of life to another in order for the firm to make appropriate and timely offers to the customer.

We develop 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 the behavior given the state. Furthermore, we extend the traditional HMM to a PHMM, which naturally fuses longitudinal (social network) activity data with one time survey data that asked 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 PHMM accurately predicts which users are active job seekers, both for out-of-sample users and out-of-time periods. Importantly, we show that the proposed model can also predict how long users are job searching and when they transition into the job seeking state. An observed state model such the ordered logit model was not able to capture such patterns. Additionally, our proposed approach allows the firm to identify which platform activities are most associated with job search and the different forms of job seeking with respect to activity on the platform. Finally, we demonstrate the marketing value of predicting the latent state by applying our proposed model 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 42%. Thus, the proposed approach offers a considerable improvement over the targeting practice observed in the data.

In this paper, we obtained rather unique and rich data 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 dataset, there are also several 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. At the same time, our goal is not to generate macro job seeking trends but rather to develop an approach that can identify active job seekers at the individual level over time. To investigate the extent of self-selection, we compare users who responded to both surveys (n=491) to users that 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-sq=2.08, P-value=0.72).

Thus, we conclude that, while self-selection may exist with respect to platform activity, it does not relate to the job seeking status. Nevertheless, our results should be particularly applicable to the somewhat more active user group. Indeed, for users with very limited activity on the platform (e.g., those who are not likely to search for a job using the social network platform), it would be very difficult to identify their job seeking status from their activity data. Future research could explore the generalizability of our approach to a more diverse and less engaged userbase.

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, Johnson and Schmittlein 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 can be seen 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 suffer from 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 from activity *prior* to validation survey, because by definition, mere-measurement effects can only occur after the measurement (the validation survey).

Third, one could argue that one may use a 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 company that provided the data and some 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 supports this notion. We find that those who were actively job searching according to their survey response in month 5, were more likely to modify their position during the three months *prior* to the survey than those who were in the passive or non-job seeking state (F-value=51.11, P-value < 0.001, N=2,814). 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, while 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 constraint in our analysis by the sample size of the survey responses from the surveys, which is relatively small compared to the userbase of most online social networking platforms. However, we note that while 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 userbase (Section 5.2).

We suggest future research to explore methods to "stochastically" fuse the survey responses with longitudinal activity data within the HMM framework. Our proposed approach treats the observed survey responses as revealing the "true" job seeking states and fuses the responses as partially observed states. One could argue that this is too restrictive and that one should allow for some slack in the data fusion, capturing measurement error in the survey responses. For instance, one could probabilistically fuse the survey responses as an additional activity variable that is only observed for some users in some time periods.

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 pregnancy, re-location, 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.

REFERENCES

Ansari, A., Montoya, R., & Netzer, O. (2012). Dynamic learning in behavioral games: A hidden Markov mixture of experts approach. *Quantitative Marketing and Economics*, *10*(4), 475-503.

Ascarza, E., & Hardie, B. G. (2013). A joint model of usage and churn in contractual settings. *Marketing Science*, *32*(4), 570-590.

Ascarza, E., Netzer, O. & Hardie, B.G. (2018). Some customers would rather leave without saying goodbye. *Marketing Science* 37(1), 54-77.

Atchadé, Y. F., & Rosenthal, J. S. (2005). On adaptive Markov Chain Monte Carlo algorithms. *Bernoulli*, *11*(5), 815-828.

Bronnenberg, B. J., Dubé, J. P. H., & Gentzkow, M. (2012). The evolution of brand preferences: Evidence from consumer migration. *American Economic Review*, *102*(6), 2472-2508.

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

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, 49, 327-335.

Ebbes, P., Grewal, R., & DeSarbo, W. S. (2010). Modeling strategic group dynamics: A hidden Markov approach. *QME*, *8*(2), 241-274.

Eddy, S.R. (1998). Profile hidden Markov models. *Bioinformatics*, 14(9), 755-763.

Feit, E. M., Beltramo, M. A., & Feinberg, F. M. (2010). Reality check: Combining choice experiments with market data to estimate the importance of product attributes. *Management Science*, 56(5), 785-800.

Feit, E. M., Wang, P., Bradlow, E. T., & Fader, P. S. (2013). Fusing aggregate and disaggregate data with an application to multi-platform media consumption. *Journal of Marketing Research*, 50 (June), 348-364.

Fong, N. M., Fang, Z., & Luo, X. (2015). Geo-conquesting: Competitive locational targeting of mobile promotions. *Journal of Marketing Research*, *52*(5), 726-735.

Ford, B.M. (1983). An overview of hot-deck procedures. In *Incomplete Data in Sample Surveys*, Vol. 2. New York: Academic Press, 185-207.

Garg, R., & Telang, R. (2017). To be or not to be linked: Online social networks and job search by unemployed workforce. *Management Science*, Articles in Advance 21 Jul 2017. https://doi.org/10.1287/mnsc.2017.2784

Gilula, Z., McCulloch, R. E., & Rossi, P. E. (2006). A direct approach to data fusion. *Journal of Marketing Research*, *43*(1), 73-83.

Granovetter, M. (1973). Weak ties and strong ties. *American Journal of Sociology*, 78, 1360-1380.

Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2), 357-384.

Hauser, J. R., Urban, G. L., Liberali, G. & Braun, M. (2009). Website morphing. *Marketing Science*, 28(2), 202-224.

Hill, K., "How Target figured out a teen girl was pregnant before her father did." *Forbes, Inc* (2012).

Kamakura, W. A., & Wedel, M. (1997). Statistical data fusion for cross-tabulation. *Journal of Marketing Research*, 485-498.

Kamakura, W. A., & Wedel, M. (2000). Factor analysis and missing data. *Journal of Marketing Research*, *37*(4), 490-498.

Leeflang, P. S. H., Wieringa, J. E., Bijmolt, T. H. A., & Pauwels, K. H. (2015). *Modeling Markets: Analyzing Marketing Phenomena and Improving Marketing Decision Making*. Springer, New York.

Matz, S. C., & Netzer, O. (2017). Using Big Data as a window into consumers' psychology. *Current Opinion in Behavioral Sciences*, *18*, 7-12.

Monaco, J. V. & Tappert, C. C. (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 Science,* 23(4), 579-595

Morwitz, V. G., Johnson, E., & Schmittlein, D. (1993). Does measuring intent change behavior?. *Journal of consumer research*, *20*(1), 46-61.

Netzer, O., Lattin, J. M., & Srinivasan, V. (2008). A hidden Markov model of customer relationship dynamics. *Marketing science*, *27*(2), 185-204.

Netzer, O., Ebbes, P., & Bijmolt, T. H. (2017). Hidden Markov Models in Marketing. In *Advanced Methods for Modeling Markets* (pp. 405-449). Springer, Cham.

Nwe, T. L., Foo, S. W., & De Silva, L. C. (2003). Speech emotion recognition using hidden Markov models. *Speech Communication*, 41(4), 603-623.

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

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77, 257-286.

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

Rossi, P. E., McCulloch, R. E., & Allenby, G. M. (1996). The value of purchase history data in target marketing. *Marketing Science*, *15*(4), 321-340.

Scheffer, T., Decomain, C. & Wrobel, S. (2001). Active Hidden Markov Models for information extraction. *Lecture Notes in Computer Science, Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis, 2189*, 309–18.

Schmidt 2017, Global Recruitment Industry Outlook for 2017 and 2018, MarketResearch.com. https://blog.marketresearch.com/global-recruitment-industry-outlook-for-2017

Schweidel, D.A., Bradlow, E.T. and Fader, P.S. (2011). Portfolio dynamics for customers of a multiservice provider. *Management Science*, 57(3), 471-486.

Schwartz, E. M., Bradlow, E. T., & Fader, P. S. (2014). Model selection using database characteristics: Developing a classification tree for longitudinal incidence data. *Marketing Science*, 33(2), 188-205.

Stopfer, J. M., & Gosling, S. D. (2013). Online social networks in the work context. In D. Derks & A. Bakker (Eds.), *The psychology of digital media at work* (pp. 39–59). London: Psychology Press.

Thompson, C. S., Thomson, P. J. & Zheng, X. (2007). Fitting a multisite daily rainfall model to New Zealand data. *Journal of Hydrology, 340*, 25–39.

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

Wedel, M., Pieters, R., & Liechty, J. (2003). Evidence for covert attention switching from eye-movements. Reply to commentaries on Liechty et al., 2003. *Psychometrika*, *68*(4), 557-562.

Wedel, M., Pieters, R., & Liechty, J. (2008). Attention switching during scene perception: how goals influence the time course of eye movements across advertisements. *Journal of Experimental Psychology: Applied*, *14*(2), 129.

Wedel, Michel, and P. K. Kannan. "Marketing analytics for data-rich environments." *Journal of Marketing* 80.6 (2016): 97-121.

Yamato, J., Ohya, J., & Ishii, K. (1992, June). Recognizing human action in time-sequential images using hidden markov model. In *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on* (pp. 379-385). IEEE.

Zarate, Luis E., Bruno M. Nogueira, Tadeu RA Santos, and Mark AJ Song. "Techniques for missing value recovering in imbalanced databases: Application in a marketing database with massive missing data." In *Systems, Man and Cybernetics, 2006. SMC'06. IEEE International Conference on*, vol. 3, pp. 2658-2664. IEEE, 2006.

Zhang, J. Z., Netzer, O., & Ansari, A. (2014). Dynamic targeted pricing in B2B relationships. *Marketing Science*, *33*(3), 317-337.

WEBAPPENDIX

USING SOCIAL NETWORK ACTIVITY DATA TO IDENTIFY AND TARGET JOB SEEKERS

This web appendix contains the following main sections:

A. Markov Chain Monte Carlo (MCMC) algorithm details for the proposed Partially Hidden Markov Model (PHMM)

B. Posterior mean and standard deviation variance-covariance matrix heterogeneity distribution for the PHMM with three states

C. Parameter estimates of the benchmark ordered logit model

D. Prediction metrics

E. PHMM estimation results for the second dataset applying the model to d-mail targeting (Section 6 in the main document)

F. Selecting the number of PHMM states for a PHMM where the number of states does not equal the number of survey response categories

G. Web appendix references

Appendix A: Markov Chain Monte Carlo (MCMC) algorithm details for the proposed Partially Hidden Markov Model (PHMM)

Given the presence of user-level heterogeneity, a Bayesian framework to estimate the PHMM is most appropriate (e.g. Netzer et al. 2017; Netzer et al. 2008; Schweidel et al. 2011; Ascarza and Hardie 2013; Zhang et al. 2014). We use a Markov Chain Monte Carlo (MCMC) algorithm to sample the posterior distribution directly through Metropolis-Hastings (MH) steps (Hastings, 1970). Given the complexity of our model due to the multivariate nature of the activity variables and the mix of distribution types, we block the parameters into separate sets and each set of parameters is updated separately using a MH step. We implement the adaptive MH algorithm as described in Atchadé and Rosenthal (2005), in short AR, which automatically adjusts the tuning parameter (the variance of the proposal density) of the MH algorithm, to improve efficiency of the MH algorithm. The proposed model presented in the main document has $K = 3$, where K is the number of states. An extension for $K > 3$ is discussed in the main document in section 7 (and Web Appendix F). We next outline the main block steps of our MCMC sampler. After that we discuss the detailed implementation of each step.

(1) Update in a MH step the parameters of the logit probabilities for the discrete activity variables δ_{0pk} and δ_{1p} for $p = 1,2,\dots,P_1$, $k = 1,2,\dots,K$, using a multivariate normal proposal density with AR flexible tuning.

(2) Update in a MH step the location parameters of the Tobit models for the continuous activity variables β_{0pk} and β_{1p} for $p = 1,2,\dots,P_2$, $k = 1,2,\dots,K$, using a multivariate normal proposal density with AR flexible tuning.

(3) Update in a MH step the variance parameters of the Tobit models for the continuous activity variables σ_{pk}^2 for $p = 1,2,\dots,P_2$, $k = 1,2,\dots,K$. Here we create proposals for log (σ_{pk}^2), which facilitates the implementation of AR flexible tuning from a multivariate normal proposal density.

(4) Update in a MH step the baseline logit threshold parameters for the initial state τ_j, $j = 1,2,\dots,K-1$ and transition probabilities ϕ_{kj} $k = 1,2,\dots,K$, $j = 1,2,\dots,K-1$ using a multivariate normal proposal density with AR flexible tuning.

(5) Update in a MH step the user-level heterogeneity parameters of the PHMM α_i for $i = 1,2,\dots,N$, using a multivariate normal proposal density with AR flexible tuning.

(6) Update in a standard Gibbs step the scale parameter Σ_α of the upper-level Normal model for the user-level heterogeneity.

Metropolis-Hastings steps (1)—(5)

As steps (1)—(5) are conceptually the same, we discuss how to generate draws for these parameters in each step of the MCMC algorithm for the general case. Let ψ denote the $R \times 1$ vector of parameters to be updated (e.g., in step (1) above, ψ would contain the elements δ_{0pk} and δ_{1p} for $p = 1,2,\dots,P_1$, $k = 1,2,\dots,K$). Let θ be the vector containing all other model parameters excluding the parameters in ψ, and let Y be the $NP \times T$ matrix of observed activity. As the full conditional distribution $f(\psi|\theta, Y)$ does not have a closed form expression for our model, we use a MH step to generate a new value for ψ in each step of the MCMC algorithm.

We generate a proposal value for ψ, say ψ_p, from a R-variate normal proposal distribution with the current value ψ, say ψ_c, as the mean, and $\tau_{MH}\Omega_{MH}$ as the variance covariance matrix, where τ_{MH} is a (scalar) parameter and Ω_{MH} is a $R \times R$ positive definite symmetric matrix. Both τ_{MH} and Ω_{MH} are flexible tuning parameters that are updated using an algorithm proposed by Atchadé and Rosenthal (2005), which we outline below.

The proposed value ψ_p is accepted with probability:

$$min\left\{\frac{\exp\left(-1/2(\psi_p-\psi_0)'V_0^{-1}(\psi_p-\psi_0)\right)L(Y|\psi_p,\theta)}{\exp(-1/2(\psi_c-\psi_0)'V_0^{-1}(\psi_c-\psi_0))L(Y|\psi_c,\theta)},1\right\},$$

where $L(Y|\psi_p,\theta)$ is the value of the full-sample likelihood, given parameters (ψ_p,θ), which is developed in the main document. Furthermore, ψ_0 and V_0 are the mean and variance-covariance matrix, respectively, of the R-variate normal prior distribution. In our study, we set $\psi_0 = 0$ and $V_0 = 100 \times I_R$ for all parameters.

The flexible tuning parameters τ_{MH} and Ω_{MH} are updated after the first 1,000 iterations. Initially, we set Ω_{MH} to the identity matrix and τ_{MH} to a value such that proposals are accepted in a broad range of 20-80%. This requires initial tuning, which we found to be easily doable. After 1,000 iterations, the parameters τ_{MH} and Ω_{MH} are automatically adjusted to target an acceptance probability of ζ, which we set to 0.28.

The main steps to automatically tune τ_{MH} and Ω_{MH} are the following (for details and proofs we refer to Atchadé and Rosenthal, 2005). Let $\epsilon_1 = 10^{-7}, \epsilon_2 = 10^{-6}, A_1 = 10^7$ and $g_{MH} = 10/l$ where l is the l-th iteration of the MCMC sampler. First, the parameter τ_{MH} is updated in the l-th iteration of the MH algorithm as:

$$\tau_{MH}^{(l+1)} = \tau_{MH}^{(l)} + g_{MH} \times (\zeta^{(l)} - \zeta) \text{ if } \epsilon_1 < \tau_{MH}^{(l+1)} < A_1,$$

$$\tau_{MH}^{(l+1)} = \epsilon_1 \text{ if } \tau_{MH}^{(l+1)} < \epsilon_1, \text{ and}$$

$$\tau_{MH}^{(l+1)} = A_1 \text{ if } \tau_{MH}^{(l+1)} > A_1.$$

Here, $\zeta^{(l)}$ is the current accept rate in the l-th iteration of the MH algorithm. In other words, if the current accept rate $\zeta^{(l)}$ is below (above) the target ζ accept rate, the updated value $\tau_{MH}^{(l+1)}$ will

be decreased (increased), which reduces (increases) the variance in the proposal distribution above. As such, the future proposed values are more (less) likely to be accepted.

The second tuning parameter Ω_{MH} is updated in the l-th iteration of the MH algorithm as:

$$\Omega_{MH}^{(l+1)} = \Gamma_{MH}^{(l+1)} + \epsilon_2 \times I_R,$$

where the parameter $\Gamma_{MH}^{(l+1)}$ is computed as:

$$\Gamma_{MH}^{(l+1)} = \Gamma_{MH}^{(l)} + g_{MH} \times \left(\left(\left(\psi_c - \mu_{MH}^{(l)} \right) \left(\psi_c - \mu_{MH}^{(l)} \right)' \right) - \Gamma_{MH}^{(l)} \right),$$

with

$$\mu_{MH}^{(l+1)} = \mu_{MH}^{(l)} + g_{MH} \times \left(\psi_c - \mu_{MH}^{(l)} \right).$$

Loosely speaking, $\mu_{MH}^{(l+1)}$ approximates the posterior mean and $\Gamma_{MH}^{(l+1)}$ approximates the posterior variance-covariance matrix of the parameter ψ when l becomes large. Let $d_{MH}^1 = \sqrt{\sum_{i,j} \left(\Gamma_{MH}^{(l+1)}(i,j) \right)^2}$, i.e., the square root of the sum of all squared elements of $\Gamma_{MH}^{(l+1)}$, then $\Gamma_{MH}^{(l+1)} = (A_1 / d_{MH}^1) \times \Gamma_{MH}^{(l)}$ if $d_{MH}^1 > A_1$. Similarly, let $d_{MH}^2 = \sqrt{\sum_i \left(\mu_{MH}^{(l+1)}(i) \right)^2}$, i.e., the square root of the sum of all squared elements of $\mu_{MH}^{(l+1)}$, then $\mu_{MH}^{(l+1)} = (A_1 / d_{MH}^2) \times \mu_{MH}^{(l)}$ if $d_{MH}^2 > A_1$. The parameters d_{MH}^1 and d_{MH}^2 prevent $\Gamma_{MH}^{(l+1)}$ and $\mu_{MH}^{(l+1)}$ from drifting away to infinity.

We note that for the variance parameters in the Tobit model we specify a log normal prior (e.g., Zellner, 1971). The advantage of such a specification in our particular case is that Atchadé and Rosenthal's (2005) algorithm for flexible tuning in the MH algorithm can be straightforwardly adapted (step 3 above). We separately update the variance parameters for the variables that we observe in each time period and the variables which we partly observe during the observation window.

Step (6) – Gibbs step to generate a draw for the scale parameter Σ_α of the upper-level Normal model for the user-level heterogeneity

The full conditional distribution or Σ_α is given by

$$p(\Sigma_\alpha|--)\sim IW(f_N, S_N^{-1}),$$

where '$--$' indicates all data and all other parameters, IW is the inverse Wishart probability density distribution, with degrees of freedom

$$f_N = f_0 + N,$$

where f_0 is the prior degrees of freedom and N is the number of users in the sample, and scale matrix

$$S_N = S_0 + \sum_{i=1}^{N} (\alpha_i - 0)\,(\alpha_i - 0)',$$

where α_i is the vector of user-level heterogeneity random intercepts of length (say) R and S_0 is the prior scale matrix. We set $S_0 = I_R$ and $f_0 = R + 15$ a priori.

Starting values of the MCMC algorithm and convergence

To facilitate faster convergence of the MCMC sampler, we first estimate a basic HMM using maximum-likelihood without user-level heterogeneity (Netzer et al. 2017). Then, we take the maximum-likelihood estimates for the parameters as starting values for the MCMC sampler, along with random starts for the heterogeneity parameters. To estimate the final model (with data fusion of the survey responses to the job seeking question) we first estimate the model without fusing the survey data and use these estimates as starting values for the PHMM. We run the MCMC chain for one million iterations burn-in, after which we retain the next 250,000 iterations

for posterior summary (to reduce computational/memory burden, we only retained every 25th iteration). Convergence was monitored by inspection of iteration plots of the sampler outputs.

Lastly, when the number of latent PHMM states exceeds the number of survey categories (in our case, when $K > 3$; See Section 7 in the main document), one needs to control for label switching for the latent states that correspond to the same survey category in estimation, by either ordering the expected values for one of the activities (e.g., Netzer et al. 2008) or by post-processing techniques (Celeux 1998). We used the post processing approach.

Appendix B: Posterior mean and standard deviation variance-covariance matrix heterogeneity distribution for the PHMM with three states

α_{i1}^{M}	2.23									
α_{i2}^{M}	1.20	0.92								
α_{i1}^{π}	-1.41	-0.99	1.30							
α_{i2}^{π}	-0.78	-0.52	0.68	0.44						
α_{i11}^{Q}	-0.43	-0.33	0.41	0.21	0.23					
α_{i12}^{Q}	-0.33	-0.31	0.44	0.27	0.16	0.42				
α_{i21}^{Q}	-1.32	-1.18	1.52	0.79	0.52	0.67	2.28			
α_{i22}^{Q}	-0.57	-0.41	0.61	0.39	0.23	0.34	0.81	0.75		
α_{i31}^{Q}	-2.61	-1.62	2.03	1.12	0.64	0.66	2.26	0.98	3.57	
α_{i32}^{Q}	-1.10	-0.79	1.03	0.59	0.36	0.44	1.27	0.70	1.65	1.06

Table B1. Posterior means lower-triangular matrix Σ_{α}. α_{i1}^{M} is the random intercept for the variable total searches and α_{i2}^{M} for the variable pageviews.

α_{i1}^{M}	(0.18)									
α_{i2}^{M}	(0.09)	(0.06)								
α_{i1}^{π}	(0.24)	(0.15)	(0.36)							
α_{i2}^{π}	(0.21)	(0.13)	(0.24)	(0.19)						
α_{i11}^{Q}	(0.16)	(0.10)	(0.16)	(0.10)	(0.11)					
α_{i12}^{Q}	(0.21)	(0.14)	(0.23)	(0.15)	(0.12)	(0.23)				
α_{i21}^{Q}	(0.22)	(0.16)	(0.35)	(0.25)	(0.21)	(0.32)	(0.64)			
α_{i22}^{Q}	(0.16)	(0.12)	(0.26)	(0.18)	(0.15)	(0.25)	(0.45)	(0.42)		
α_{i31}^{Q}	(0.33)	(0.21)	(0.45)	(0.34)	(0.23)	(0.33)	(0.50)	(0.39)	(0.81)	
α_{i32}^{Q}	(0.18)	(0.11)	(0.24)	(0.19)	(0.15)	(0.21)	(0.30)	(0.28)	(0.36)	(0.27)

Table B2. Posterior standard deviations lower-triangular matrix Σ_{α}. α_{i1}^{M} is the random intercept for the variable total searches and α_{i2}^{M} for the variable pageviews.

Appendix C: Parameter estimates of the benchmark ordered logit model

Parameter		B	Std. Error	Wald Chi-Square	Sig.
Threshold	[Active seekers]	-2.556	.7616	11.260	.001
	[Passive seekers]	.424	.7476	.322	.571
Total searches		.090	.1181	.585	.444
Pageviews		-.072	.1150	.389	.533
Profile updates (dum)		-.071	.3017	.055	.814
Job searched (dum)		-1.056	.3469	9.260	.002
Invitations sent		-.255	.1907	1.786	.181
Invitations received		.360	.2378	2.297	.130
Connections formed		-.007	.2622	.001	.980
Number of connections of invitee		.097	.0671	2.082	.149
More invitations outside company (dum)		-.472	.6504	.526	.468
Total searches – lag		.001	.1366	.000	.993
Pageviews – lag		-.183	.1123	2.657	.103
Profile updates (dum) – lag		.784	.3396	5.324	.021
Job searched (dum) – lag		-.326	.3947	.681	.409
Invitations sent – lag		.177	.2150	.681	.409
Invitations received – lag		-.081	.2417	.111	.739
Connections formed – lag		.105	.2737	.146	.702
Number of connections of invitee – lag		-.122	.0684	3.181	.074
More invitations outside company (dum) – lag		-.199	.6348	.098	.754

Table C1. Ordered logit model for predicting job seeking status in month 5. The dependent variable is an ordinal variable capturing the Survey 1 responses to whether user is active, passive, or not job seeking in month 5 ($N = 400$).

We note that for predicting the job seeking status in month 14 with the ordered logit model, we do not observe the covariates Invitations sent, Invitations received, Connections formed, Number of connections of invitee, and More invitations outside company (dum), as explained in the main document. We therefore re-estimated the ordered logit model without these covariates included before generating the predictions for month 14.

Parameter		B	Std. Error	Wald Chi-Square	Sig.
Threshold	[Active seekers]	-2.009	.2586	60.355	.000
	[Passive seekers]	.905	.2349	14.857	.000
Total searches		.072	.1142	.397	.529
Pageviews		-.107	.0934	1.305	.253
Profile updates (dum)		-.059	.2937	.041	.840
Job searched (dum)		-1.118	.3405	10.780	.001
Total searches – lag		.031	.1306	.057	.811
Pageviews – lag		-.104	.0864	1.451	.228
Profile updates (dum) – lag		.843	.3287	6.581	.010
Job searched (dum) – lag		-.367	.3884	.892	.345

Table C2. Ordered logit model for predicting job seeking status in month 14. The dependent variable is an ordinal variable capturing the survey 1 responses to whether user is active, passive, or not job seeking in month 5 ($N = 400$).

Appendix D: Prediction metrics

		Observed in survey	
		Non/passive job seeker (0)	Active job seeker (1)
Prediction	Non/passive job seeker (0)	C00	C01
	Active job seeker (1)	C10	C11

Table D1. Table of prediction classification counts. Here C00 is the number of correctly predicted non/passive job seekers, C01 is the number of active job seekers that were not predicted by the model (false negatives), C10 is the number of non/passive job seekers that were predicted by the model as being active seekers (false positives) and C11 is the number of correctly predicted active job seekers.

In Table D1, C11 are the true positives (TP), C10 are the false positives (FP), C01 are the false negatives (FN) and C00 are the true negatives (TN). Using the notation of Table D1, the three metrics used in the main document are computed as follows. The Jaccard Similarity Index is defined as $J = \frac{C11}{C11+C10+C01}$. The Jaccard index measures the "share" of the correct active job seeking predictions. It is maximal (=1) when FP and FN are zero. The Fowlkes-Mallows (FM) Index is given by $FM = \sqrt{\frac{C11}{C11+C10} \times \frac{C11}{C11+C01}}$. It is maximal when the false predictions C10 and C01 are both 0. Lastly, the Classification Success Index is defined as: $CSI = 1 - \left(1 - \frac{C11}{C11+C10} + 1 - \frac{C11}{C11+C01}\right) = \frac{C11}{C11+C10} + \frac{C11}{C11+C01} - 1$, where $1 - \frac{C11}{C11+C10}$ correspond to the Type 1 error and $1 - \frac{C11}{C11+C01}$ to the Type 2 error of predicting active job seekers. The CSI index captures the measure of minimal error and is maximal when both errors are minimal. It ranges from –1 (both errors are maximal) to 1 (both errors are minimal). We note that the value zero has no particular meaning.

Appendix E: PHMM estimation results for the second dataset applying the model to d-mail targeting (Section 6 in the main document)

We obtained a second sample of 1,621 users from the platform. For this sample we observe the users' activity for a different sample of users than the ones used in Section 5, during the time period June 2011—May 2012. These users responded to the same job search survey as discussed in Section 3 of the main document. This survey was fielded in October 2011 (5^{th} month of the observation window). For these users we also observe whether they received an d-mail in each month and, if they received an d-mail, whether they responded to it. The activity variables we observe for this sample are similar, but not identical, to the activity variables observed in the sample used for the main analyses. We observe the following monthly activities: whether or not the user viewed a job ad (0/1 variable), whether or not the user updated his/her educational information on the profile page (0/1 variable), whether or not the user updated his/her position information on the profile page (0/1 variable), how many invitations to connect (s)he received, how many invitations to connect (s)he sent, how many page views the user made, and how many times the user's profile page was viewed. All variables are observed for the full time period.

The posterior results for the proposed PHMM are given in table E1. As in the main text, we report the transformed working parameters (α_i and θ). The trend parameters are reported at the working parameter level. Similar to the results in the main document, we masked the absolute monthly activity levels by multiplying them with the same random number, which was a single draw from a uniform distribution on the interval [0.5, 1.5].

State	Non Seeking	Passive	Active	Trend parameter
Job views (dummy)	0.01 (0.00)	0.15 (0.01)	0.50 (0.01)	0.007 (0.003)
Education modified (dummy)	0.01 (0.00)	0.04 (0.00)	0.21 (0.01)	-0.041 (0.005)
Positions modified (dummy)	0.02 (0.00)	0.11 (0.01)	0.44 (0.01)	-0.032 (0.003)
Invitations received	1.88 (0.04)	2.66 (0.03)	3.88 (0.07)	0.008 (0.001)
Invitations sent	3.24 (0.26)	3.00 (0.08)	18.84 (1.07)	-0.019 (0.002)
Page views	21.39 (0.86)	214.02 (4.20)	952.60 (35.75)	0.008 (0.001)
Profile views received	2.96 (0.13)	9.18 (0.21)	44.31 (1.61)	-0.004 (0.001)
Initial state distribution	0.37 (0.01)	0.42 (0.01)	0.21 (0.01)	
Transition matrix				
	0.45 (0.01)	0.39 (0.01)	0.16 (0.01)	
	0.21 (0.01)	0.58 (0.01)	0.21 (0.01)	
	0.15 (0.01)	0.40 (0.02)	0.44 (0.01)	

Table E1. Posterior means and standard deviations (in parentheses) for the proposed PHMM.

When we compare the posterior estimates of the PHMM for this dataset with the posterior estimates of the three state model in the main document (Table 2), we can see that the findings are fairly similar. The active job seekers exhibit the highest activity, followed by passive job seekers and non-seekers. In addition, the posterior results for the initial state distribution and the transition probability matrix are very similar to the posterior results reported in the main document (Table 2). Thus, these findings also suggests that the insights reported in the main document are fairly robust, and generalizable to a new sample of users.

Appendix F: Selecting the number of PHMM states for a PHMM where the number of states does not equal the number of survey response categories

In this appendix we propose an approach to select the number of states in a PHMM, when the number of states is larger than the number of survey response categories. In such a case, we need to both select the number of states and the mapping between the PHMM states and the survey response categories. Bayesian model fit criteria such as the log marginal density (LMD) often tend to under-penalize complex models with a large number of states (Netzer et al. 2017). Furthermore, our interest is not to predict user activity on the social network platform but rather to predict the nature of the latent job seeking states. Accordingly, we contrast the LMD, which measures fit with respect to the site activity variables with predictions of the latent variable - job seeking states.

To do so, we split our calibration sample ($N = 400$; Section 5 main document) into two samples ($N_1 = 300$ and $N_2 = 100$). We fit the model on data from the 300 users and predict job seeking status in month 5 for the remaining 100 users. Specifically, we compute the same three metrics (Jaccard index (J), the Fowlkes–Mallows index (FM), and the Classification success index (CSI)) of job seeking prediction in the validation sample (N_2), as discussed in Section 5.2 of the main document and Web Appendix D. We use the model predictions and the job seeking status reported in the first survey in month 5 to calculate the three prediction metrics. We use the same procedure described in Section 5.2 of the main document to obtain the individual level parameters (α_i) for the holdout sample users ($N_2 = 100$).

We consider the three state, all four and five state models and some of the six state models (the choices for six state models were guided by our findings for the four and five state models). The LMD and three validation metrics results are reported in Table F1.

	LMD	J	FM	CSI
NPA	**-19,896**	**0.171**	**0.305**	**-0.362**
NNPA	-19,413	0.172	0.294	-0.412
NPPA	**-19,347**	**0.174**	0.307	-0.365
NPAA	-19,480	0.173	**0.329**	**-0.266**
NNNPA	-19,134	0.154	0.275	-0.433
NNPPA	-18,846	0.074	0.140	-0.716
NNPAA	-19,003	0.111	0.211	-0.554
NPPPA	**-18,748**	0.125	0.229	-0.527
NPPAA	-18,832	0.158	0.280	-0.425
NPAAA	-19,059	**0.194**	**0.409**	**0.026**
NNPPPA	-18,550	0.074	0.140	-0.716
NNPPAA	**-18,477**	**0.176**	**0.303**	**-0.386**
NPPPAA	-18,507	0.156	0.271	-0.456
NPPAAA	-18,599	0.140	0.268	-0.413

Table F1: LMD ($N_1 = 300$) and validation metrics ($N_2 = 100$) for selecting the number of states and the state mapping. Bold figures represent best model within the number of states.

From Table F1 we can see that the overall best model to recover whether a user is a job seeker is the five state NPAAA model. This model has the highest Jaccard, FM, and CSI values. The LMD measure keeps increasing as we increase the number of states. Increasing the number of states to six states does not seem to improve the ability to predict job seekers. For the four and five states model it appears that models that map the observed active job seeking status from the survey into multiple PHMM states (the NPAA and NPAAA models) perform best in terms of predicting active job seekers.

We contrast these two models (NPAA and NPAAA) with the NPA and the ordered logit benchmark model in terms of holdout sample predictions both out-of-sample and out-of-time. Table F2 extends Table 5 in the main document. Considering the results in Table F2, it follows that the prediction ability of the three PHMMs (NPA, NPAA, NPAAA) is fairly similar, and all three PHMMs outperform the ordered logit benchmark model in predicting active job seekers.

			Time							
			Month 5 – Survey 1				Month 14 – Survey 2			
			NPA	NPAA	NPAAA	Ord. Logit	NPA	NPAA	NPAAA	Ord. Logit
Cross section	Calibration sample ($N_c = 400$)	J	N/A				0.21	0.19	0.20	0.15
		FM					0.35	0.34	0.36	0.28
		CSI					-0.30	-0.27	-0.21	-0.40
	Holdout sample ($N_h = 91$)	J	0.28	0.26	0.20	0.26	0.23	0.26	0.24	0.15
		FM	0.45	0.45	0.39	0.44	0.37	0.42	0.40	0.29
		CSI	-0.08	-0.02	-0.10	-0.08	-0.26	-0.14	-0.17	-0.33

Table F2. Results holdout predictions for PHMM: NPA, NPAA, and NPAAA and the ordered logit model. Performance metrics (J, FM, and CSI) indicate model performance to predict whether a user is an active job seeker in month 5 and month 14. Higher numbers indicate better performance.

Appendix G: Web appendix references

Celeux, G. (1998). Bayesian inference for mixture: The label switching problem.
In *Compstat* (pp. 227-232). Physica, Heidelberg.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1), 97-109.

Zellner, A. (1971). *An introduction to Bayesian inference in econometrics*. J. Wiley and Sons, Inc.