A Sequential Approach to Market State Modeling and Analysis in Online P2P Lending

Hongke Zhao, *Member, IEEE*, Qi Liu, Hengshu Zhu, Yong Ge, Enhong Chen, *Senior Member, IEEE*, Yan Zhu, and Junping Du

***Abstract*—Online peer-to-peer (P2P) lending is an emerging wealth-management service for individuals, which allows lenders to directly bid and invest on the listings created by borrowers without going through any traditional financial intermediaries. As a nonbank financial platform, online P2P lending tends to have both high volatility and liquidity. Therefore, it is of significant importance to discern the hidden market states of the listings (e.g., hot and cold), which open venues for enhancing business analytics and investment decision making. However, the problem of market state modeling remains pretty open due to many technical and domain challenges, such as the dynamic and sequential characteristics of listings. To that end, in this paper, we present a focused study on market state modeling and analysis for online P2P lending. Specifically, we first propose two enhanced sequential models by extending the Bayesian hidden Markov model (BHMM), namely listing-BHMM (L-BHMM) and listing and marketing-BHMM (LM-BHMM), for learning the latent semantics between listings' market states and lenders' bidding behaviors. Particularly, L-BHMM is a straightforward model that only considers the local observations of a listing itself, while LM-BHMM considers not only the listing information but also the global information of current market (e.g., the competitive and complementary relations among listings). Furthermore, we demonstrate several motivating applications enabled by our models, such as bidding prediction and herding detection. Finally, we construct extensive experiments on two real-world data sets and make some deep analysis on bidding behaviors, which clearly validate the effectiveness of our models in terms of different applications and also reveal some interesting business findings.**

***Index Terms*—Bayesian hidden Markov model (BHMM), bidding behaviors, market state, peer-to-peer lending.**

I. Introduction

RECENT years have witnessed the rapid development and prevalence of online peer-to-peer (P2P) lending platforms, such as Prosper[1] and LendingClub.[2] As an emerging wealth-management service for individuals, P2P lending allows individuals to directly borrow and lend money from one to another without going through traditional financial intermediaries. Indeed, P2P lending has become a fast growing investment market with more than 100% year over year growth.[3] For instance, LendingClub announced its total loan issuance amount had reached more than US $13.4 billion at the end of 2015.

In P2P lending market, there are two main kinds of roles: 1) the *borrowers* who want to borrow money from others and 2) the *lenders* who lend money to borrowers. Trading in this market follows the *Dutch Auction Rule* [1], [2]. Specifically, for borrowing money, the borrower will first create a *listing* to solicit bids from lenders by describing herself and the reason of borrowing money (e.g., for wedding). Then, if a lender wishes to invest on this listing within its soliciting duration (e.g., one week), a bid is created by her describing how much money she wants to invest (e.g., $50) and the minimum rate. Finally, the listings which can receive enough money in time will turn to loans and begin the repayment periods. Otherwise, all the previous investments on these listings will be canceled.

In the literature, the rapid prevalence of P2P lending has triggered many important research problems, such as listing quality or borrower credit evaluation [3]–[5], bidding behaviors analyzing [6]–[8], and loan funding prediction [9], [10]. Unfortunately, the problem of modeling market states of listings (e.g., hot and cold) is still under-explored. Indeed, discerning market state is critically important due to the high volatility and liquidity of online P2P lending. At the micro level, market state modeling can help borrowers monitor the listing popularity and present better listings. Meanwhile, it also can help lenders understand the listings more clearly and make better investment decision, e.g., quick investment, avoiding

Manuscript received October 21, 2015; revised September 3, 2016; accepted January 21, 2017. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1000904, in part by the National Science Foundation for Distinguished Young Scholars of China under Grant 61325010, and in part by the National Natural Science Foundation of China under Grant U1605251, Grant 61672483, and Grant 61532006. The work of Q. Liu was support by the Youth Innovation Promotion Association of CAS under Grant 2014299. This paper was recommended by Associate Editor F. Wang. *(Corresponding author: Qi Liu.)*

H. Zhao, Q. Liu, E. Chen, and Y. Zhu are with the School of Computer Science and Technology, University of Science and Technology of China, Hefei 230026, China (e-mail: zhhk@mail.ustc.edu.cn; qiliuql@ustc.edu.cn; cheneh@ustc.edu.cn; zhuyan90@mail.ustc.edu.cn).

H. Zhu is with Big Data Laboratory, Baidu Research, Beijing 100085, China (e-mail: zhuhengshu@baidu.com).

Y. Ge is with the Eller College of Management, University of Arizona, Tucson, AZ 85721 USA (e-mail: yongge@email.arizona.edu).

J. Du is with the School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: junpingd@bupt.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMC.2017.2665038

[1]https://www.prosper.com/

[2]https://www.lendingclub.com/

[3]http://techcrunch.com/2012/05/29/peer-to-peer-lending-crosses-1-billion-in-loans-issued/

bidding failure, *herding* [7], [11], [12], or serious *competition* [7]. Moreover, at the macro level, market state modeling can be used for enhancing market management, such as listing recommendation and listing shelves guidance. However, there are many technical and domain challenges on market state modeling. First, how to capture the dynamic and sequential characteristics of market state is not trivial. Second, how to construct and mine the relevance between the lenders' bidding behaviors and the hidden market states is another open question. At last, how to capture the market information/situation for a listing is also worth exploring, e.g., although a listing was hot in the previous periods, it may suffer from the emerging competition when the market changes and then turn to a cold state.

To address the above challenges, in this paper, we present a focused study on market state modeling in P2P lending by extending Bayesian hidden Markov model (BHMM) with different domain assumptions. Specifically, we first model the dynamic and sequential characteristic of market state using the Markov chain structure. Here, we assume that the market state of a listing is influenced by its own properties, and its current state is only determined by its previous one state. Based on this assumption, we propose a listing-BHMM (L-BHMM) model for modeling market state. Also, real-world findings reveal that market state may be influenced by the market situation. Therefore, we further design a more comprehensive model named listing and marketing-BHMM (LM-BHMM) for the holistic consideration. In this way, both L-BHMM and LM-BHMM could model the listings by connecting the hidden market states of listings and the bidding behaviors of lenders. Furthermore, we demonstrate two motivating financial applications enabled by market state modeling, namely bidding prediction and herding detection. Finally, extensive experiments based on two real-world data sets clearly validate the effectiveness of our models in terms of different applications, and reveal some interesting business findings on bidding behaviors. To the best of our knowledge, this is the first comprehensive study on market state modeling in online P2P lending. This paper, we hope, will be helpful for enhancing P2P lending services, such as bidding prediction, investment decision making, and market management.

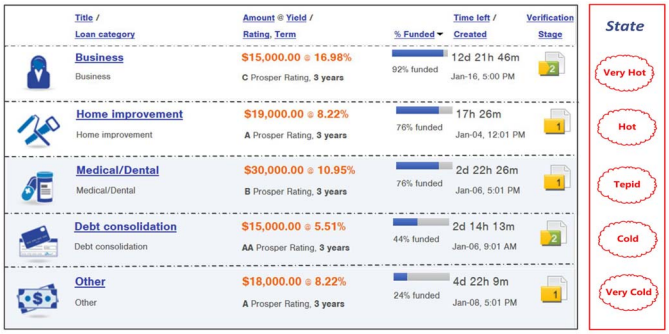
Overview: The remainder of this paper is organized as follows. In Section II, we formulate the research problem and introduce the construction of observations from bidding behaviors and market information. In Section III, we show the details of our proposed models (L-BHMM and LM-BHMM) and the applications of market state modeling. Section IV presents the experimental results and some findings from deep analysis. Finally, we briefly introduce the related works and conclude our work in Sections V and VI.

II. Preliminaries

In this section, we first introduce the research problem of market state modeling and then present the details of constructing observations from lenders' bidding behaviors and market information. For facilitate illustration, Table I lists the mathematical notations used in this paper.

TABLE I
Mathematical Notation

Notation	Description
S_t	a listing's market state with timestamp t
O_t	the bid observation with timestamp t
M_t	the market observation with timestamp t
$\theta_{S_{t-1}}$	market state transition distribution of $(S_t \vert S_{t-1})$ in L-BHMM
θ_{S_{t-1},M_t}	market state transition distribution of $(S_t \vert S_{t-1}, M_t)$ in LM-BHMM
ϕ_{S_t}	output emission distribution of the bid observation on market state S_t
α, β	the Dirichlet priors for Θ, Φ
J	the number of unique bid observations
K	the number of market states
H	the number of market observations
R	all the listing records/time slices
L	all the listings



(a) (b)

Fig. 1. (a) Listings in P2P lending. (b) Market states (manually labeled) of corresponding listings.

A. Problem Statement

In P2P lending, a listing is the predecessor of a loan, only if this listing could receive enough money within the soliciting duration (e.g., one week). Fig. 1(a) shows an illustration of several real listings from Prosper, one of the largest P2P lending platforms in America. From this figure, we can obtain some basic properties and descriptions of a listing that are given by the borrower, e.g., the loan category/purpose (e.g., Business), the required amount (e.g., $15 000.00), and the rate (e.g., 16.98%). Meanwhile, we can observe some dynamic information of a listing from the incremental lenders' bidding behaviors, such as the current funded amount (e.g., 92% funded) and the remaining time (e.g., 12 day 21 h 46 min). One step further, the local dynamic information of each individual listing will make up the global information of the entire market. From these observations in Fig. 1, we could learn that the first two listings are much more popular (hotter) than the last one. As the first two listings have received most of the required bids with a long remaining time while the last one has only received small amount of bids with a short remaining time. That is, if properly modeled, this observed information could reflect the hidden market state of a listing. For example, the current probable hidden market states (manually labeled) of the listings are shown in Fig. 1(b). In summary, in this paper, we aim to automatically detect and model such hidden market states for listings through the observations of lenders' bidding behaviors and the market situation. Meanwhile, we will further apply the learned market states for developing some important financial applications. Here, we formally define the problem of market state modeling as follows.

Definition 1 (Market State Modeling): Given a set of listings L, in which each listing $l \in L$ has a sequence of historical observations, e.g., bid observations and market observations (defined in Section II-B). The problem of market state modeling is to learn a model from all the observation sequences, which can be used for predicting or detecting the listings' market states or observations in the future.

B. Observation Construction

As the above description, the market state is dynamic and sequential, i.e., the market state of a listing may change every day in its soliciting duration and correspondingly we may observe different observations in different timestamp. Thus, in this section, we introduce how to construct granulometric observations for modeling the hidden market states of listings, i.e., construct bid observation by leveraging the bidding behaviors, and construct the market observation using the holistic market information.

1) Bid Observation: In this paper, the lenders' bidding behaviors on listings are one of the most important dynamic information, which reflect the lenders' current acceptance degree to these listings. During the auctions, listings with hotter states will receive more bids while colder listings receive fewer bids. Thus, we can adopt the average bidding amount of each period (e.g., one day) as the main observation. However, for different listings, the length of the entire duration for soliciting bids and the totally required amount are also different. Thus, it is appealing to define and construct the comparable bid observations for all the listings. In this paper we define the bid observation as follows.

Definition 2 (Bid Observation): A bid observation O_t in period t of a listing with totally required amount A and entire soliciting duration P, is equal to $O_t = \Delta \times (P/A)$, where Δ is the average bidding amount during period t.

In the above definition, we use P/A for eliminating the effects of required amounts and soliciting durations of different listings. In this way, the bid observations are comparable for different listings, and they could be applied to model the market state of listings.

2) Market Observation: Besides the lenders' bidding information, the holistic market information is also very helpful for market state modeling. For instance, a high-risk listing which suffers from a *cold* state may transfer to a *hot* state when the market situation is beneficial for it, e.g., many complementary (low-risk) listings appear in the market. In contrast, it may stay cold if there are too many similar/competitive (high-risk) listings at that time [7]. This kind of influence from market information to the listing states does exist because of the investment composition effect in economic field or P2P lending. Specifically, the rational lenders have the *portfolio perspective* [13] in their minds, and they would like to optimize their portfolio by investing several listings with different risk values. As a result, the complementary listings of a listing can promote the acceptance degree of lenders to this listing, and vice versa. According to existing studies in the fields of economics and biology, the relations of goods (i.e., listings in this paper) in a market are mainly determined by their similarities [14], [15]. Thus, we construct the market observation as follows.

Definition 3 (Market Observation): A market observation M_t in period t of a listing l is defined as the average similarity between l and other listings in the market in period t. Specifically, $M_t = \mathrm{Sim}(l, \neg l) = (1/|\neg l|) \sum_{l' \in \neg l} \mathrm{Cos}(l, l')$, where $\neg l$ are the listings in their soliciting bid durations at time t except listing l.

In mature P2P lending platforms, e.g., Prosper, hundreds of listings are soliciting bids at the same time every day, so that it is pointless to consider all the other listings when calculating the market observations for a specific listing. In fact, for a given listing l, the most influential listings to l are those with nearest auction durations. Thus, in practice, we use the nearest neighbor (e.g., 20) listings with similar start time to construct set $\neg l$ for l. In the following, we show how to compute the cosine similarity between two listings, i.e., $\mathrm{Cos}(l, l')$. In P2P lending market, *return* and *risk* are two most important aspects for assessing a listing [16], [17]. Thus, motivated by the method in [17], the listing profile can be represented as a two-element vector $\overrightarrow{P_l} = [P_l, R_l]$, where the first term is the expected return which is a given value and the second term is the estimated risk. Here, we use the lend rate declared by borrower in the soliciting duration of a listing as the expected return term, and meanwhile, adopt a logistic regression model to estimate the risk value for each listing. The logistic regression model is widely used in risk assessment due to its simplicity and relatively high performance [3], [17]. More detailed information about the computation of P_l and R_l, which are the same with those in [16] and [17], are omitted due to the limited space. Finally, the cosine similarity of two listings in the market observation can be calculated by

$$\mathrm{Cos}\big(l, l'\big) = \frac{P_l P_{l'} + R_l R_{l'}}{\sqrt{P_l^2 + P_{l'}^2}\sqrt{R_l^2 + R_{l'}^2}}. \tag{1}$$

Based on Definitions 2 and 3, the variables of bid observations and market observations for each listing in any periods could be computed, and then, they will be used for modeling the listings' market states.

3) Observation Segmentation: According to the definitions of bid observation (O_t) and market observation (M_t), these observations are continuous numerical variables. If we directly take the original observations into model training, the convergence may be very difficult to guarantee since the granularity of observations is too fine and chaotic. Thus, we propose an information gain based method [18], [19] to preprocess the original continuous numerical variables into discrete interval variables/segments. Here, we take the segmentation of bid observations as an example, and the market observations could be processed in the same way. Specifically, the information entropy of the observation interval/segment O^S is given by

$$\mathrm{Ent}\Big(O^S\Big) = -\sum_{i=1}^{|O^S|} p_i \log(p_i) \tag{2}$$

where $|O^S|$ is the number of observation subsegments in segment O^S, and p_i is the proportion of subsegment O^{P_i}, i.e., number of observations in this subsegment divided by the total number of observations in O^S. The segment process is conducted as follows. First, we rank the observations of all the

listings in each period (e.g., in every day) based on their original values. Second, all the sorted observations are viewed as a big initial segment and we partition it into several subsegments in a recursive binary way. In each iteration, we use the *weighted average entropy* (denoted as WAE) to find the best split position

$$\mathrm{WAE}\left(i; O^S\right) = \frac{\left|O_1^S(i)\right|}{\left|O^S\right|}\mathrm{Ent}\left(O_1^S(i)\right) + \frac{\left|O_2^S(i)\right|}{\left|O^S\right|}\mathrm{Ent}\left(O_2^S(i)\right) \quad (3)$$

where $O_1^S(i)$ and $O_2^S(i)$ are two subsegments of segment O^S when being split at the ith observation. The best split position indicates a maximum information gain given by $\Delta E(i)$ which is equal to $\mathrm{Ent}(O^S) - \mathrm{WAE}(i; O^S)$.

After the segment process, original bid observations and market observations of all the listings in each period are processed into discrete intervals according to the observation values. Now, we can take these segment intervals as our new observable variables. Without loss of generality, the observations mentioned in the rest of this paper are all the discrete intervals after segment.

III. Market State Modeling

In this section, we present the details of our approach to market state modeling. Specifically, we first propose two specific models to connect the hidden states of listings and the bidding behaviors of lenders. Then, we introduce two applications enabled by market state modeling, namely, bidding prediction and herding detection.

A. L-BHMM Model

Here, we propose the first model named L-BHMM to model the market state in P2P lending. The basic assumption about the state dependency in L-BHMM is as follows.

Assumption 1: The current market state of a listing is only determined by the previous one state of this listing.

The above assumption indeed follows the first-order Markov property, which indicates that the market state of a listing only depends on its previous lenders' acceptance degree. This assumption is straightforward and easy to think of. Indeed, the first-order Markov property has been referred and demonstrated in previous relevant works, e.g., stock market state [20], [21], recruitment market [22], and even collective evolution inference in P2P lending network [23]. Following this assumption, we apply the first-order Markov chain structure to listing state modeling. Further, as a variation of the classical Hidden Markov Model, the robustness and scalability of BHMM in modeling sequential data with external knowledge have been well proved in previous studies [24], [25]. To that end, we also propose to leverage BHMM for modeling market state by considering the sequential and dependency characteristics of bidding behavior information. Specifically, L-BHMM has the same structure of a standard bi-gram Hidden Markov Model which contains Dirichlet priors over transition and emission distributions for modeling the sequential listing records (i.e., bid observation). Fig. 2 shows the graphical representation of L-BHMM model, where S_t is the market state in period t, and O_t is the bid observation in period t. Please note that, we assume that state transition and observation emission of all listings share the same distributions Θ and Φ, respectively.

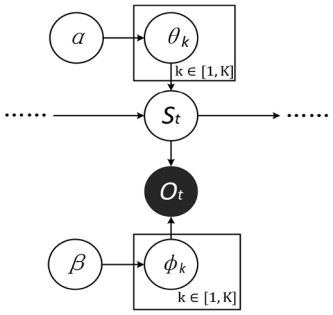


Fig. 2. Graphical representation of L-BHMM.

The parameterizations of L-BHMM are as follows:

$$\begin{aligned} S_t|S_{t-1}, \Theta &\sim \mathrm{Mult}\left(\theta_{S_{t-1}}\right) \\ O_t|S_t, \Phi &\sim \mathrm{Mult}\left(\phi_{S_t}\right) \\ \theta_{S_{t-1}}|\alpha &\sim \mathrm{Dirichlet}(\alpha) \\ \phi_{S_t}|\beta &\sim \mathrm{Dirichlet}(\beta) \end{aligned}$$

where $S_t|S_{t-1}, \Theta \sim \mathrm{Mult}(\theta_{S_{t-1}})$ means the current state S_t follows multinomial distribution $\mathrm{Mult}(\theta_{S_{t-1}})$ given the previous state S_{t-1} and Θ; and $O_t|S_t, \Phi \sim \mathrm{Mult}(\phi_{S_t})$ means the current bid observation O_t follows multinomial distribution $\mathrm{Mult}(\phi_{S_t})$ given the current state S_t and Φ. Θ is the state transition distribution and Φ is the output emission distribution of the bid observations on the market state. Particularly, both Θ and Φ follow the Dirichlet distributions with parameter α and parameter β.

The generative process of L-BHMM is as follows. First, a prior transition distribution of listing's market state (i.e., θ) is generated from a prior Dirichlet distribution α. Similarly, a prior output distribution of bid observations (ϕ) is generated from a prior Dirichlet distribution β. Second, a listing state S_t is generated from distribution θ_{t-1} with respect to the previous listing state S_{t-1}. Finally, a bid observation O_t is generated from distribution ϕ_{S_t} with respect to the listing's current market state S_t.

Given the hyperparameters α and β in this generative model, we can calculate the joint distribution of L-BHMM by the following:

$$\begin{aligned} &P(O_t, S_t, \Theta, \Phi|S_{t-1}, \alpha, \beta) \\ &= P(\Theta|\alpha)P(\Phi|\beta)P(S_t|S_{t-1}, \Theta)P(O_t|S_t, \Phi) \\ &= P(\Theta|\alpha)P(\Phi|\beta)P\left(S_t|S_{t-1}, \theta_{S_{t-1}}\right)P\left(O_t|S_t, \phi_{S_t}\right). \end{aligned} \quad (4)$$

Therefore, the likelihood of a set of records R can be calculated as follows[4]:

$$\begin{aligned} L(R) = &\int \prod_{k=1}^{K} P(\theta_k|\alpha) \prod_{t=1}^{|R|} P\left(S_t|S_{t-1}, \theta_{S_{t-1}}\right) d\Theta \\ &\times \int \prod_{k=1}^{K} P(\phi_k|\beta) \prod_{t=1}^{|R|} P\left(O_t|S_t, \phi_{S_t}\right) d\Phi. \end{aligned} \quad (5)$$

[4] R represents all the listings. For convenient writing, we use a same notation t to represent the sequences of all listings.

The objective of training L-BMMM is to learn the proper hidden variables Θ and Φ to maximize the likelihood function in (5). However, the likelihood function is complicated which is not easy to calculate directly. In this paper, we adopt the Gibbs sampling method [26], [27], a form of Markov chain Monte Carlo, to estimate the parameters. Given parameters α and β, the training process begins with a random assignment of market states to all bid observations for initializing the state of Markov chain. In each iteration, the method will re-estimate the conditional probability by assigning a market state to each bid observation, which is conditioned by the assignment of all other observations except the current one. Then, according to the conditional probabilities, a new assignment of observation to listing state will be updated as a new market state of the Markov chain. Finally, after enough iterations, the assignment will converge and every bid observation is assigned to a stable market state.

According to the Gibbs sampling rule, we need to calculate the conditional distribution $P(S_t = k|\boldsymbol{S}_{\neg \boldsymbol{t}}, R)$, where $\boldsymbol{S}_{\neg \boldsymbol{t}}$ means the market states for all bid observations except O_t. It should be noted that, in L-BHMM, records R only contain the bid observation sequence $\boldsymbol{O} = (O_1, O_2, \ldots)$. The parameter of state transition distribution Θ is $K \times K$ dimension and output emission distribution Φ is $K \times J$ dimension. We can derive the conditional distribution as follows:

$$\begin{aligned}
&P(S_t|\boldsymbol{S}_{\neg \boldsymbol{t}}, R) \propto P(S_t, \boldsymbol{S}_{\neg \boldsymbol{t}}, R)\\
&= P(R|S_t, \boldsymbol{S}_{\neg \boldsymbol{t}})P(S_t|\boldsymbol{S}_{\neg \boldsymbol{t}})P(\boldsymbol{S}_{\neg \boldsymbol{t}})\\
&= P(R|S_t, \boldsymbol{S}_{\neg \boldsymbol{t}})P(S_t|\boldsymbol{S}_{\neg \boldsymbol{t}})P\big(S_{t+1}, \boldsymbol{S}_{\neg (\boldsymbol{t},\boldsymbol{t+1})}\big)\\
&= P(R|S_t, \boldsymbol{S}_{\neg \boldsymbol{t}})P(S_t|\boldsymbol{S}_{\neg \boldsymbol{t}})P\big(S_{t+1}|\boldsymbol{S}_{\neg (\boldsymbol{t},\boldsymbol{t+1})}\big)P\big(\boldsymbol{S}_{\neg (\boldsymbol{t},\boldsymbol{t+1})}\big)\\
&\propto P(R|S_t, \boldsymbol{S}_{\neg \boldsymbol{t}})P(S_t|\boldsymbol{S}_{\neg \boldsymbol{t}})P\big(S_{t+1}|\boldsymbol{S}_{\neg (\boldsymbol{t},\boldsymbol{t+1})}\big)\\
&= P(O_t|S_t, \boldsymbol{S}_{\neg \boldsymbol{t}}, \boldsymbol{\Lambda}_{\neg \boldsymbol{t}})P(\boldsymbol{\Lambda}_{\neg \boldsymbol{t}}|\boldsymbol{S_t}, \boldsymbol{S}_{\neg \boldsymbol{t}})P(S_t|\boldsymbol{S}_{\neg \boldsymbol{t}})\\
&\quad \times P\big(S_{t+1}|\boldsymbol{S}_{\neg (\boldsymbol{t},\boldsymbol{t+1})}\big)\\
&= P(O_t|S_t, \boldsymbol{S}_{\neg \boldsymbol{t}}, \boldsymbol{\Lambda}_{\neg \boldsymbol{t}})P\big(\boldsymbol{\Lambda}_{\neg \boldsymbol{t}}|\boldsymbol{S}_{\neg \boldsymbol{t}}\big)P(S_t|\boldsymbol{S}_{\neg \boldsymbol{t}})\\
&\quad \times P\big(S_{t+1}|\boldsymbol{S}_{\neg (\boldsymbol{t},\boldsymbol{t+1})}\big)\\
&\propto P(O_t|S_t, \boldsymbol{S}_{\neg \boldsymbol{t}}, \boldsymbol{\Lambda}_{\neg \boldsymbol{t}})P(S_t|\boldsymbol{S}_{\neg \boldsymbol{t}})P\big(S_{t+1}|\boldsymbol{S}_{\neg (\boldsymbol{t},\boldsymbol{t+1})}\big)
\end{aligned}$$

where $\boldsymbol{\Lambda}$ is the set of all the bid observations in R, and $\boldsymbol{\Lambda}_{\neg \boldsymbol{t}}$ means removing bid observation O_t from R. We calculate the three multipliers in (6), respectively, and the final result are as follows[5]:

$$\begin{aligned}
&P(S_t = k|\boldsymbol{S}_{\neg \boldsymbol{t}}, R)\\
&\propto \frac{n_{\neg \boldsymbol{t},S_t,O_t} + \beta_{O_t}}{\sum_{j=1}^{J} n_{\neg \boldsymbol{t},S_t,j} + \beta_j} \times \frac{n_{\neg \boldsymbol{t},(S_{t-1},S_t)} + \alpha_{S_t}}{\sum_{k'=1}^{K} n_{\neg \boldsymbol{t},(S_{t-1},k')} + \alpha_{k'}}\\
&\quad \times \frac{n_{\neg (\boldsymbol{t},\boldsymbol{t+1}),(k,S_{t+1})} + \boldsymbol{I}(S_{t-1} = k = S_{t+1}) + \alpha_{S_t}}{n_{\neg (\boldsymbol{t},\boldsymbol{t+1}),k} + \boldsymbol{I}(S_{t-1} = k) + \sum_{k'=1}^{K} \alpha_{k'}}
\end{aligned} \tag{6}$$

where $n_{\neg \boldsymbol{t},S_t,j}$ is the number of the jth bid observation in bid records except the tth record with assignment market state k, $n_{\neg \boldsymbol{t},(S_{t-1},k')}$ is the number of the previous states are S_{t-1} and current states are assigned to k' in all records except the tth record, $n_{\neg (\boldsymbol{t},\boldsymbol{t+1}),(k,S_{t+1})}$ is the number of current states are assigned to k and the next states are S_{t+1} in all records except in the tth record and $(t+1)$th record, and $\boldsymbol{I}(\boldsymbol{x})$ is an indicator function whose value is 1 when x is true and 0 otherwise.

After enough rounds of iteration, the assignment will converge, thus, we can estimate the parameters Θ and Φ by

$$\theta_{S_{t-1},k} = \frac{n_{S_{t-1},k} + \alpha_k}{\sum_{k'} n_{S_{t-1},k'} + \alpha_{k'}},\ \phi_{S_t,j} = \frac{n_{S_t,j} + \beta_j}{\sum_{j'} n_{S_t,j'} + \beta_{j'}} \tag{7}$$

where $n_{S_{t-1},k}$ is the number of current market states that are assigned to k and the previous states are assigned to S_{t-1}, $n_{S_t,j}$ is the number of current state that equals S_t and the bid observation is j.

In summary, in L-BHMM, we only take use of the bid observation to estimate the transitions of market state (i.e., Θ) and the observation emission (i.e., Φ). In the next section, we will introduce another more comprehensive model which considers both the bid information and the market information for listing's market state modeling.

B. LM-BHMM Model

As illustrated in Assumption 1, L-BHMM indicates that each market state is only dependent on the previous acceptance degree from lenders to this specific listing. That is, the listings with hot state are more likely to be hot and listings with cold state are more likely to be cold in the future. However, in P2P lending, the market state of one listing may be also influenced by the market situation, e.g., a listing which suffers from a cold state may transfer to a hot state when many complementary listings appear in the market. Thus, in this section, we propose a more reasonable assumption for the state dependency of listings. Then, based on the proposed assumption, we show the details of the proposed LM-BHMM for market state modeling.

Assumption 2: The current market state of a listing is determined by both its previous state and the current market situation.

In other words, for modeling the market state of a given listing, Assumption 1 only considers the previous observation of this listing while Assumption 2 also considers the global observations of some other listings in a given period of the market. Based on Assumption 2, we design the LM-BHMM for market state modeling. Specifically, Fig. 3 shows the graphical representation of LM-BHMM. We can see that, compared to Fig. 2, one more observed variable M_t, which denotes the market observation in period t, also influences the listing market state S_t. LM-BHMM is more comprehensive and an extension of L-BHMM.

Similarly, the parameterizations of LM-BHMM are

$$\begin{aligned}
S_t|S_{t-1}, M_t, \Theta &\sim \text{Mult}\big(\theta_{S_{t-1},M_t}\big)\\
O_t|S_t, \Phi &\sim \text{Mult}\big(\phi_{S_t}\big)\\
\theta_{S_{t-1},M_t}|\alpha &\sim \text{Dirichlet}(\alpha)\\
\phi_{S_t}|\beta &\sim \text{Dirichlet}(\beta)
\end{aligned}$$

where $S_t|S_{t-1}, M_t, \Theta \sim \text{Mult}(\theta_{S_{t-1},M_t})$ means the current state S_t follows multinomial distribution $\text{Mult}(\theta_{S_{t-1},M_t})$ given previous state S_{t-1}, current market observation M_t and state transition distribution Θ. Please note that, different from L-BHMM, the parameter of the state transition distribution

[5]We omit the detailed derivation due to the limited space.

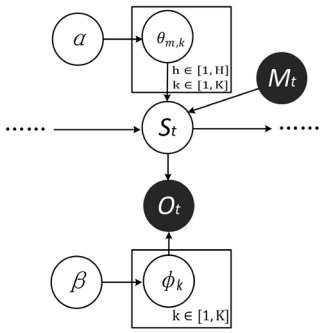


Fig. 3. Graphical representation of LM-BHMM.

Θ is $K \times K \times H$ dimension in LM-BHMM, i.e., the current state of a listing is determined by both the previous state and current market situation observation of this listing. Thus, in the training of LM-BHMM, the record set R contains not only the bid observations $\boldsymbol{O} = (O_1, O_2, \ldots)$ but also the market observations $\boldsymbol{M} = (M_1, M_2, \ldots)$. The generative process of LM-BHMM is similar to that of L-BHMM except for the state transition. Specifically, state S_t is generated from θ_{S_{t-1},M_t} with respect to the previous state S_{t-1} and current market observation M_t.

Given the hyperparameters α, β in this generative model, we can calculate the joint distribution of LM-BHMM by the following:

$$\begin{aligned} &P(O_t, S_t, M_t, \Theta, \Phi|S_{t-1}, \alpha, \beta) \\ &= P(\Theta|\alpha)P(\Phi|\beta)P(S_t|S_{t-1}, M_t, \Theta)P(O_t|S_t, \Phi) \\ &= P(\Theta|\alpha)P(\Phi|\beta)P\big(S_t|S_{t-1}, M_t, \theta_{S_{t-1},M_t}\big)P\big(O_t|S_t, \phi_{S_t}\big). \end{aligned} \tag{8}$$

Thus, the likelihood of a set of records R can be calculated

$$\begin{aligned} L(R) = &\int \prod_{h=1}^{H}\prod_{k=1}^{K} P\big(\theta_{h,k}|\alpha\big) \prod_{t=1}^{|R|} P\big(S_t|S_{t-1}, M_t, \theta_{S_{t-1},M_t}\big)d\Theta \\ &\times \int \prod_{k=1}^{K} P(\phi_k|\beta) \prod_{t=1}^{|R|} P\big(O_t|S_t, \phi_{S_t}\big)d\Phi. \end{aligned} \tag{9}$$

Similar to (6), we can calculate the conditional distribution $P(S_t = k|\boldsymbol{S}_{\neg t}, R)$ of LM-BHMM as follows:

$$\begin{aligned} &P(S_t|\boldsymbol{S}_{\neg t}, R) \\ &\propto P(O_t|S_t, \boldsymbol{S}_{\neg t}, \boldsymbol{\Lambda}_{\neg t})P(S_t|\boldsymbol{S}_{\neg t}, M_t)P\big(S_{t+1}|\boldsymbol{S}_{\neg(t,t+1)}\big). \end{aligned} \tag{10}$$

One step further, it could be represented by

$$\begin{aligned} &P(S_t = k|\boldsymbol{S}_{\neg t}, R) \\ &\propto \frac{n_{\neg t,S_t,O_t} + \beta_{O_t}}{\sum_{j=1}^{J} n_{\neg t,S_t,j} + \beta_j} \times \frac{n_{\neg t,(S_{t-1},M_t,S_t)} + \alpha_{S_t}}{\sum_{k'=1}^{K} n_{\neg t,(S_{t-1},M_t,k')} + \alpha_{k'}} \\ &\times \frac{n_{\neg(t,t+1),(k,S_{t+1})} + \boldsymbol{I}(S_{t-1} = k = S_{t+1}) + \alpha_{S_t}}{n_{\neg(t,t+1),k} + \boldsymbol{I}(S_{t-1} = k) + \sum_{k'=1}^{K} \alpha_{k'}}. \end{aligned} \tag{11}$$

Indeed, the main difference between (6) and (11) lies in the second multiplier, where $n_{\neg t,(S_{t-1},M_t,S_t)}$ is the number of market states that are assigned to state S_t when the corresponding market observation is M_t and previous listing states are assigned to state S_{t-1} in all the records R except for the tth one. Finally, the observation emission Φ is computed the same as those in L-BHMM, while the transitions of listing state Θ can be estimated as follows:

$$\theta_{(S_{t-1},M_t),k} = \frac{n_{(S_{t-1},M_t),k}}{\sum_{k'=1}^{K} n_{(S_{t-1},M_t),k'} + \alpha_{k'}}. \tag{12}$$

Indeed, in L-BHMM and LM-BHMM, we leverage the construction of Markov Chain to connect the observable variables (i.e., bid observation and market observation) extracted from lenders' bidding behaviors to the hidden market states of listings. Thus, our models can be applied to some applications related to the analysis of lenders' bidding behaviors.

C. Applications

In this section, we present two specific motivating applications enabled by market state modeling, namely bidding prediction and herding detection.

1) Bid Observations Prediction: One straightforward application of market state modeling for listings are to predict the future bid observations in P2P lending market. Specifically, for a target listing η, we have observed a sequence of bid observations $\boldsymbol{O}^\eta = (O_1^\eta, \ldots, O_t^\eta)$ and a sequence of market observations $\boldsymbol{M}^\eta = (M_1^\eta, \ldots, M_t^\eta)$. We can estimate the market state for each bid observation $O_i^\eta (0 \le i \le t)$ by

$$P\big(S_i^\eta|\boldsymbol{O}^\eta, \boldsymbol{M}^\eta, \Theta, \Phi\big) \propto P\big(O_i^\eta|S_i^\eta, \Phi\big)P\big(S_i^\eta|M_i^\eta, \Theta\big) \tag{13}$$

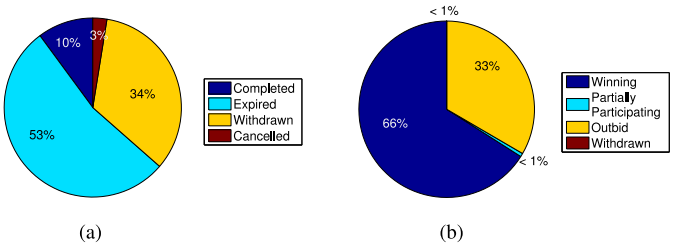
which can be computed by the *Viterbi Algorithm* [28]. Then, we could predict the (t+1)th bid observation as follows:

$$P\big(O_{t+1}^\alpha = O|S_t^\alpha, \Phi\big) = \sum_{S_{t+1}^\alpha = S'} P\big(S'|S_t^\alpha, M_{t+1}^\alpha, \Theta\big)P\big(O|S', \Phi\big). \tag{14}$$

Particularly, for the observations that does not have previous ones, we set the transition probabilities equally. By modeling the listing's market state, we can predict the future bid observations for a listing. One step further, some other specific applications could also be implemented. Here, we introduce a specific application which is related to the herding phenomenon in P2P lending [7], [12].

2) Herding Detection: In P2P lending market, when lenders make their investment decisions on listings, they may bid together on a few listings without rational investment decisions, but just follow the herd and ignore most of the remaining listings. Thus, some popular listings often receive massive biddings while some other listings receive scarce biddings. This is called as the herding phenomenon [7], [11], [12]. Although this phenomenon has been widely observed, the scientific literature on this topic is still limited, and even a clear definition of herding is lacking.

Generally, we think of two ways to detect herding. In the first way, we could view herding as a phenomenon that the bidding amount received in a period is much more than a specific threshold value Υ. For instance, we can define the threshold as $\Upsilon = \lambda \cdot (A/P)$, where A is the totally required amount of this specific listing, P is its entire soliciting duration, and λ is a manually given parameter. However, it is not easy to manually define the value of threshold. Thus, we adopt another way of herding detection. Specifically, we first estimate the



(a) (b)

Fig. 4. Percents of listings and bids in different categories. (a) Listings versus fund results. (b) Bids versus bidding results.

market states and bid observations of listings for a following period of time [i.e., by (14)]. Then, we rank all the listings based on their estimated bid observation values, and select top *Num* listings as the most suspicious listings that may involve in herding phenomenon.

IV. Experiments

In this section, we evaluate our models on two large-scale real-word data sets and make some deep analysis about bidding behaviors from the experimental results.

A. Experimental Data

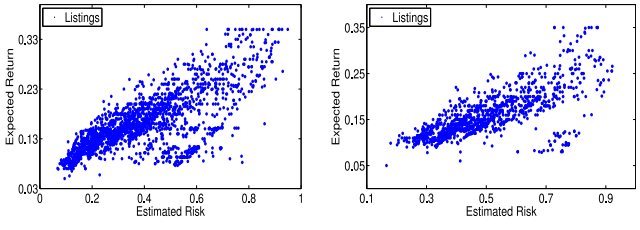
The original data set was downloaded from Prosper.com,[6] which contains all the records in this platform from November, 2005 to the end of May, 2011. Readers can also find the data from this URL.[7] From this data set, we selected three tables that are relevant to market state modeling. Specifically, *Bid* table contains the timestamp and the amount of money that lenders bid for each listing. These records are the basis used to construct bid observations. *Listing* table contains the basic properties or descriptions of listings. *LoanPerformance* table records the performances (i.e., repay in time or default) of listings. The records in *Listing* and *LoanPerformance* tables are used to train the risk prediction model, and then construct listing profiles and market observations (i.e., Section II-B).

Fig. 4 shows the percent of listings with different fund results and the percent of bids with different bidding results. In Fig. 4(a), only about 10% of listings (*completed*) could receive enough bids in the soliciting durations, and about 53% of listings fail to be fund in time (*expired*) and 34% of listings are *withdrawn* by the borrowers and 3% of listings are *canceled* by the Propose system for some reasons. In Fig. 4(b), only about 66% of bids will succeed (*Winning*) and about 33% of bids fail, i.e., they are outbid by others. Intuitively, lots of listings receive excessive bids or even cause herding phenomenon. In contrast, some other listings are not appealing enough to be invested on and they will fail to turn into loans. From the above statistical results we could summarize that the bidding performance are quite different among listings, and it is necessary to model market state for listings and better understand the lenders' bidding behaviors.

Considering the listing profiles vary a lot, we extracted two subdata sets from the original data for experiments, where the listings in each data set have the similar soliciting duration and amount of required money. Table II shows the basic statistics of these two data sets. The soliciting durations of these listings are 7 days, 10 days, or 14 days. The required amount of the listings in the first data set is around \$10 000 and the required amount of the listings in the second data set is in the range of [\$20 000, \$25 000]. Actually, the two data sets contain more than 51.5% bid records of the original data set.

TABLE II
Statistics of Experimental Data Sets

NO.	Listing Num	Required Amount($)	BidRecords Num
1	2,774	9,000-11,000	608,216
2	1,469	20,000-25,000	664,843



(a) (b)

Fig. 5. Scatter plot of listing profiles. (a) Data set 1. (b) Data set 2.

B. Experimental Setup

In experiments, we adopt fivefold cross-validation, and in each round, we randomly sampled 20% of data from each data set for test and the remaining 80% were used for training.[8] Without loss of generality, we set the bid observation segment number as 32 (i.e., $J = 32$), the market observation segment number as 4 (i.e., $H = 4$), and the number of hidden market states as 16 (i.e., $K = 16$) for both L-BHMM and LM-BHMM. The time granularity is set to 1 day for each period, i.e., we record the observations in each day. The initial probabilities are set as $1/K$ for all states.

C. Experimental Results

In the next, we will report the experimental results. Specifically, we demonstrate: 1) results and findings in the preprocess; 2) the bid prediction performance of our approaches; 3) efficiency results on model training and test processes; 4) case studies that exploit LM-BHMM on herding detection and state modeling; and 5) some deep analysis results of bid observation and application on macroscopic market.

1) Preprocess Results: Here, we first introduce the results of observation construction. For bid observations, we obtain the bidding amount for each listing everyday and then run the segment process. For market observations, we first profile each listing with respect to *return* and *risk*, then calculate the original market observation value based on the cosine similarity and run the segment process at last.

[6]https://www.prosper.com/tools/DataExport.aspx
[7]http://home.ustc.edu.cn/%7Ezhhk/DataSets.htm

[8]Please note that, the sequentiality is for each listing not for all listings. Thus, we can construct cross-validation via different listing selections.

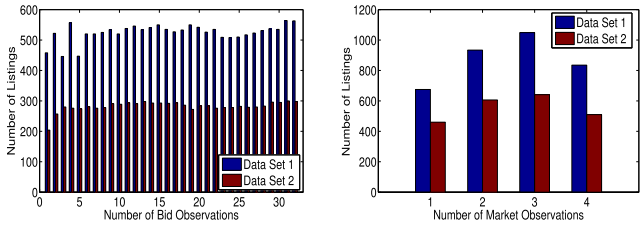


Fig. 6. Distribution of observations. (a) Bid observation. (b) Market observation.

Fig. 5 shows the scatter plot of randomly sampled listing profiles with regard to their returns and risks. We can see that the positive correlations (i.e., Pearson correlation coefficients of two data sets are: 0.733 and 0.711)[9] between each listing's expected return and estimated risk. These results are consistent with the reported results in [17]. We can also observe that listing profiles are different, e.g., some listings have low risks and low returns while some have high risks and high returns. Note that, the listings with similar profiles are competitive and the different listings are complementary. Fig. 6 shows the distribution of number of listings with respect to different observations after segment process. From these two subfigures we can see that the segment process can divide the original numerical variables into interval variables evenly for both bid observation and market observation.

2) Effectiveness of Prediction: In this section, we validate the effectiveness of our models in terms of predicting bid observations. For LM-BHMM and L-BHMM models, we adopt the famous *Viterbi Algorithm* [28] for prediction. Specifically, for each test sequence, we randomly select the first N ($N \in [1, \text{len} - 1]$, where len is the sequence length of the test listing) observation records for fitting models, and use the (N+1)th bid observation as the ground truth for evaluation.

Baselines: To the best of our knowledge, there are no existing works on market state modeling in P2P lending. We exploit two baselines in this part of experiment. The first baseline is denoted as Pre, which takes the previous bid observation as the prediction result. The second baseline is *Logistic regression* (denoted as Reg), which is widely used in time series predictions [29]. Reg takes input the first N observations and output the (N+1)th bid observations.

Metrics: We adopt two famous metrics in the information retrieval field to evaluate the accuracy of prediction results. The first metric is the hit rate (HR) [30] and the second is the root mean square error (RMSE). Their definitions are as follows:

$$\text{HR} = \frac{\sum_{i=1}^{\text{NT}} \boldsymbol{I}(p_i = t_i)}{\text{NT}},$$
$$\text{RMSE} = \sqrt{\sum_{i=1}^{\text{NT}} (p_i - t_i)^2 / \text{NT}} \tag{15}$$

where NT is the number of test sequences, p_i is the predicted value of the (N+1)th bid observation in the ith test sequence and t_i is the true value of the (N+1)th test bid observation in this sequence. $\boldsymbol{I}(\boldsymbol{x})$ is an indicator function whose value is 1 when x is true and 0 otherwise.

[9]We used more listing properties (e.g., isHomeOwner, bankAccount) than those used in [17] to learn the risk prediction model.

TABLE III
RUNNING TIME (S)

Data Set	Models	TrainingTime(Iterations)				Predict Time
		20	40	60	80	
1	Reg	0.83	1.69	2.52	3.33	0.0007
	L-BHMM	11.11	22.47	33.88	44.75	0.0177
	LM-BHMM	18.90	38.43	57.99	77.60	0.0198
2	Reg	0.43	0.87	1.30	1.76	0.0004
	L-BHMM	6.16	12.61	19.16	25.98	0.0082
	LM-BHMM	10.55	21.48	31.44	41.40	0.0089

The results of cross-validation are shown in Fig. 7. Specifically, Fig. 7(a) and (b) are the results under HR metric, which are the larger the better. From the HR results, we can see that LM-BHMM obtains the largest HRs (i.e., outperforms about 0.02-0.03 than L-BHMM and Reg) on both data sets, while L-BHMM and Reg perform similarly under HR metric. Fig. 7(c) and (d) are the results under RMSE metric, which are the smaller the better. From these results, we can see that LM-BHMM can obtain the smallest RMSE values (i.e., outperforms about 0.3–0.5 than L-BHMM and Reg) on both the two data sets. Also, on the second data set, L-BHMM model performs better than Reg. These results clearly validate the prediction effectiveness of our models, especially the sophisticated model LM-BHMM.

In the above experiments, we input the first N observation records for fitting models and use the (N+1)th bid observation as ground truth for evaluation. Indeed, it is also necessary to evaluate the effect of sequence length on prediction performances of different models. For better comparison, in this part of experiment, we only compare the performances of three models on the listings with the same sequence length (7 days), and the results are shown in Fig. 8. Here, all the results are the average HR and RMSE values obtained by five-fold cross-validation. First, we can see that the LM-BHMM performs best all the time which demonstrates the prediction performance and robustness of LM-BHMM. Second, these figures show the overall prediction accuracy trends, i.e., the HR values increase and the RMSE values decrease with the growing of the input sequence length. The results indicate that the longer sequence length contains more information on the listing dynamics, and thus, the models could capture the sequence dependence of states more accurately. However, please note that, an interesting finding is that HR values decrease and RMSE values increase when the length reaches to the largest value (i.e., the sequence length is 6). In other words, it is more difficult to predict the bidding in the end of soliciting duration. The reason for this finding will be explained in the section of bid observation analysis.

3) Efficiency Results: In experiments, all the methods are performed on the same platform, and we record the average running times of the five-fold cross-validation for three models in Table III. We can see that Reg is much faster than other models, and LM-BHMM will take more times (about 0.3/0.2 s per iteration on Data Set 1/ Data Set 2) than L-BHMM for training. However, the differences of L-BHMM and LM-BHMM in test phases are not significant.

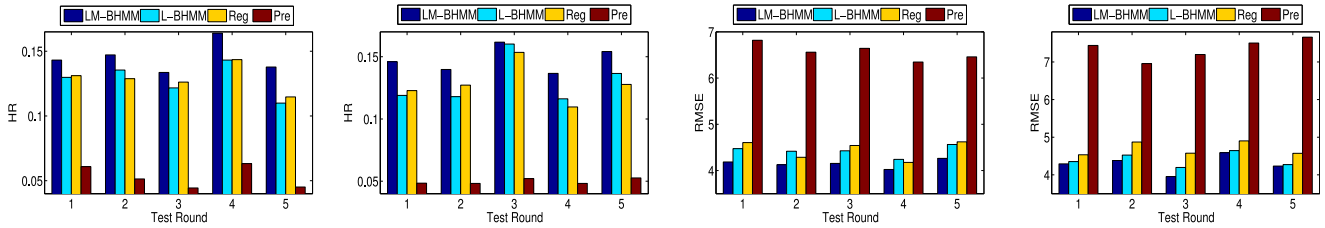


Fig. 7. Performance of bid prediction. (a) HR Results:Data Set 1. (b) HR Results:Data Set 2. (c) RMSE:Data Set 1. (d) RMSE:Data Set 2.

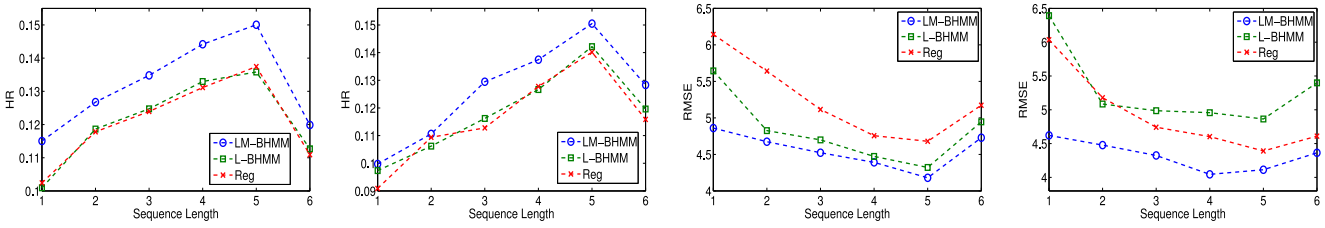


Fig. 8. Performance with respect to different sequence length. (a) HR Results:Data Set 1. (b) HR Results:Data Set 2. (c) RMSE Results:Data Set 1. (d) RMSE Results:Data Set 2.

4) Case Study of Herding Detection: In this section, we design an experimental case study of herding detection. Specifically, we first sample a small test data set (50 listings) in which listings have the same required amount ($A =$ \$10 000) and soliciting duration ($P = 7$ days). Then, we adopt LM-BHMM model to estimate the last market state and the bid observation of each listing. Further, we rank the listings based on their bid observation estimations. Fig. 9(a) reports the six representative listings in the ranked listings[10] Listings 1 and 2 are the two most suspicious listings selected from the top position of the ranked listings. Listings 3 and 4 are two listings selected from the middle position. These two listings could receive enough bids in time, meanwhile, they did not cause excessive competitions and herding phenomenon. Listings 5 and 6 are selected from the end of the ranked listings, so they are listings of colder states. These two listings can not receive enough bids in time. From this simple case study, we can see LM-BHMM can detect the herding phenomenon and distinguish different states effectively.

We also display the modeled market states for listings (Listing 1, Listing 2, Listing 3, and Listing 6) in Fig. 9(b). Since the states are hidden, we distinguish their semantics by experts. We can see that Listings 1 and 2 are very hot especially at the end of their durations. While Listing 3, especially Listing 6, is much colder.

5) Bid Observation Analysis: In the following, we will make some deep analysis about observations and lenders' bidding behaviors.

a) Bid observation distribution: First, we study the bid observation distribution with respect to the soliciting duration. We select all the listings with same soliciting durations, i.e., 7 days, in the two data sets. We label and rank all the

[10]For better exhibit, we plot three lines when the bidding amount values equal to $3 \cdot (A/P)$, $2 \cdot (A/P)$ and $1 \cdot (A/P)$, respectively.

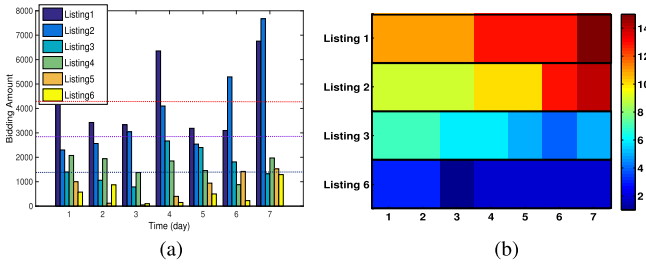


Fig. 9. Case study. (a) Herding detection. (b) State modeling.

bid observations, and divide all the bid observations into four intervals (namely: "VeryCold," "Cold," "Hot," and "VeryHot") based on the listings' rankings. Then, we take the statistics of the bid observations in the four intervals in every timestamp (day). Fig. 10 reports the statistics results on two data sets. On the whole, the proportion of VeryHot interval increases and the proportion of VeryCold interval decreases over time. From these two figures, we can infer that most listings suffer from "start-up" or "cold-start" [31] processes at the beginning phases of the whole durations while herding is more likely to occur at the ending phases of the whole durations. This phenomenon may be due to lenders' distrust to the new listings and the lenders' psychology of following suit at the ending phases. In comparison, the proportions of the middle two intervals are more stable.

Besides, a special discovery in Fig. 10 is that the proportions of VeryHot and Hot intervals in the first days are much bigger than those in the following days. That is because for many lenders, especially blind lenders, they are curious about new listings especially on the first days of soliciting durations. But over time, the lenders' bidding behaviors will stabilize gradually. Similar findings were also reported in [7] and [32].

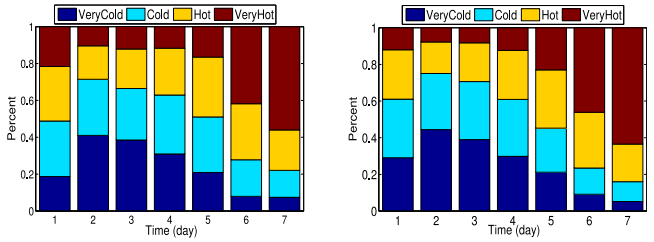


Fig. 10. Bid observation distribution versus soliciting bids time. (a) Data set 1. (b) Data set 2.

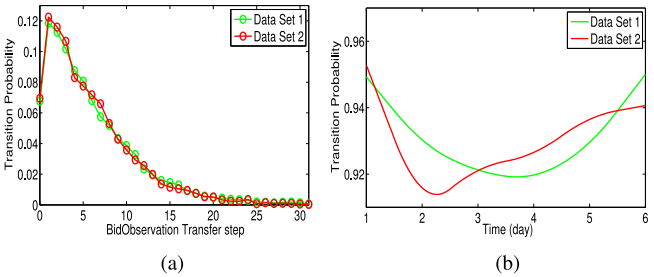


Fig. 11. Bid observation transition probability versus (a) transition step and (b) soliciting bids time.

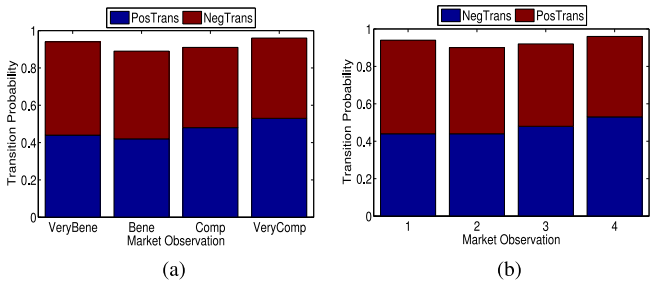


Fig. 12. Bid observation transition probability versus market observation. (a) Data set 1. (b) Data set 2.

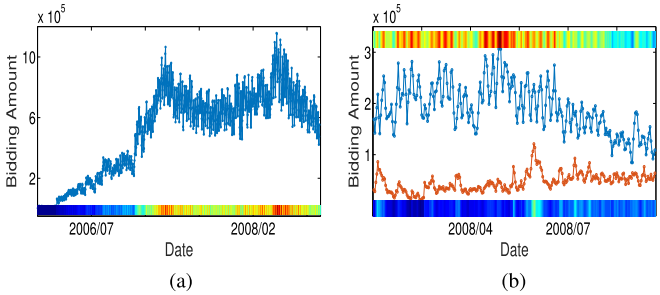


Fig. 13. Macroscopic analysis. (a) Macroscopic market. (b) Business and HomeImprovement.

b) Bid observation transition: Next, we study the transitions among bid observations. Fig. 11(a) plots the transition probability of bid observation with respect to the transition step. For instance, step equals to 0 means the state does not transfer to other states and step equals to 1 means the state transfers to its nearest neighbor states. From the overall tendency, a state is more likely to transfer to itself and its near neighbor states and the probability of state transition decreases with the increase of transition step.

In Fig. 11(b), we plot the transition probability of bid observations with respect to the time slice. From this figure we can see that, the transition probabilities of bid observations are even higher than 90%, which means predicting the bid observation is a difficult problem. What is more, the bid observations are more likely to transfer at the beginning or at the end of the entire soliciting duration, which is consistent with the results in [7]. This is because rational lenders are more likely to bid at the middle of the soliciting bid duration while the blind lenders bid more randomly. This phenomenon may be the reason why listing is difficult to predict at the beginning and end of the duration. Thus, the prediction performance will be bad when the length reaches to the largest value (Please recall the results in Fig. 8).

c) Bid observation transition under market observation: Finally, we plot the bid observation transition directions, e.g., PosTrans (if a observation transfer to another hotter observations) and NegTrans (if a observation transfer to another colder observations) probabilities under different market observations in Fig. 12. Specifically, there are four market observations, namely "VeryBenefic" (VeryBene), "Benefic" (Bene), "Competitive" (Comp), and "VeryCompetitive" (VeryComp) ranked as the predefined market situations in Section II-B2. From these two figures, we can see that transitions are more likely to occur under the VeryBene and VeryComp market situations. More specifically, when the market is VeryBene, the bid observations are more likely to transfer to the hotter states, and vice versa. These findings demonstrate the rationality of our second assumption.

d) Macroscopic analysis: In this paper, we mainly focuses on the microscopic market, i.e., the listing-level state modeling. However, our approach (i.e., L-BHMM)[11] can also be applied to the macroscopic market analysis. We obtain the global bidding for all the listings (or a certain category of listings) to construct the macroscopic bid observation. Fig. 13(a) shows the global bidding amount in Prosper from November 2005 to October 2008, in which the color bar visualizes the modeled hidden macroscopic state (i.e., the warmer the color/tone, the hotter the state). Fig. 13(b) shows the bidding and states of two categories (i.e., "Business" and "HomeImprovement") of listings from December 2007 to October 2008.

From these two figures, we can get some interesting findings for the macroscopic market. Specifically, from the long-term view, the macroscopic market is more and more popular, and attracting more users; different kinds of listings/submarkets have different market states. For example, on the whole, Business listings are much popular/hotter and changing over time. However, the HomeImprovement market is always in cold states.

V. Related Work

In general, the related works can be mainly grouped into two categories.

[11]LM-BHMM does not work any more without market observation in macroscopic view.

The first category is about P2P lending. Readers can refer to [6] and [33] for an overview of P2P lending. In this field, many studies aimed at the listing quality evaluation. For instance, Dong *et al.* [3] proposed a kind of logistic regression model with random coefficients to improve the prediction accuracy of credit scorecards. Wang *et al.* [5] proposed a fuzzy support vector machine to discriminate good creditors from bad ones. Luo *et al.* [4] developed a lender composition model to measure the listings. These works all tried to build a finer model to access the quality of listings or borrowers, thus helping lenders to make better decisions with these global models. Some other research issues in P2P lending include borrower decision optimization [34], determinants analysis of loan funding [9], [10], and raising dynamic of loans [23], [35]. For instance, Herrero-Lopez [36] measured the influence of social interactions in the risk evaluation of a money request. The results in this paper showed that fostering social features increases the chances of getting a loan fully funded, when financial features were not enough to construct a differentiating successful credit request. Freedman and Jin [37] studied whether social networks help alleviate the information problems. Besides, Zhao *et al.* [2] and [17], respectively, proposed methods for loan recommendation and portfolio selection in P2P lending market. Recently, Ceyhan *et al.* [7] made some efforts on bidding behavior analysis and observed the herding phenomenon on bidding. However, this paper was mainly from the statistical and empirical perspectives to understand lender behaviors, and lacking in the deeper explorations. In all, there are few existing works on state modeling for listings in the P2P lending market.

The second category is about Hidden Markov Models (HMMs), which has been widely used in some domains, such as signal processing and speech recognition [38], [39], economics [40], [41], and biometrics [42], [43]. Until today, Markov model is still with strong vitality in many fields. In [44], Hidden Markov Model (HMM) was adapted to developing a continuous flow model for market demand-driven systems. Zhang *et al.* [45] designed two efficient Markov chain dynamics under the data-driven Markov chain Monte Carlo framework to effectively explore the high-dimensional state space for human pose estimation. In [46], HMM model was adapted to modeling the popularity of mobile App.

For much of the history, HMMs have been implemented by recursive algorithms for parameter estimation [47]. In recent literatures, some researchers proposed using Bayesian method to estimate the parameters of HMMs, which can provide a more stable parameter estimation. For example, Goldwater and Griffiths [24] proposed a novel B-HMM model for part-of-speech tagging. Liechty *et al.* [48] developed a psychometric model of visual covert attention by B-HMM, and they tested it using eye-movement data. Guha *et al.* [49] modeled the array comparative genomic hybridization data by B-HMM model. In [25], BHMM was adapted to context recognition for mobile users. Gao and Johnson [50] compared a variety of different Bayesian estimators for HMM part-of-speech taggers with various numbers of hidden states on data sets of different sizes. Compared with the traditional HMM, previous studies have demonstrated the robustness and scalability of B-HMM in modeling sequential data with domain priori [24], [51]. However, to the best of our knowledge, neither HMM nor B-HMM has been adopted in the domain of P2P lending.

VI. Conclusion

In this paper, we proposed a focused study on market state modeling for listings in online P2P lending and designed two sequential models (i.e., L-BHMM and LM-BHMM) by extending BHMM. Specifically, the first model L-BHMM only considers the local information and observations of a listing itself, while the second model LM-BHMM considers not only the listing information but also the global information of current market. Both of these two models could reveal the latent semantics between lenders' bidding behaviors and the market states of listings. Then, we demonstrated that market state modeling can be applied to many novel applications, such as bidding prediction and herding detection. Finally, we conducted extensive experiments on two real-world data sets, and the experimental results clearly validated the effectiveness of our models. Meanwhile, on the basis of our studies, we also made some deep analysis about observations and lenders' bidding behaviors and reported some interesting findings.

In the future, we would like to explore more factors to further improve the performance of LM-BHMM. Moreover, we also plan to study more novel applications in P2P lending enabled by our market state modeling approaches.

References

[1] M. Kumar and S. I. Feldman, "Internet auctions," in *Proc. 3rd USENIX Workshop EC*, vol. 3. Boston, MA, USA, 1998, pp. 49–60.

[2] H. Zhao, Q. Liu, G. Wang, Y. Ge, and E. Chen, "Portfolio selections in P2P lending: A multi-objective perspective," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, San Francisco, CA, USA, 2016, pp. 2075–2084.

[3] G. Dong, K. K. Lai, and J. Yen, "Credit scorecard based on logistic regression with random coefficients," *Procedia Comput. Sci.*, vol. 1, no. 1, pp. 2463–2468, May 2010.

[4] C. Luo, H. Xiong, W. Zhou, Y. Guo, and G. Deng, "Enhancing investment decisions in P2P lending: An investor composition perspective," in *Proc. ACM SIGKDD*, San Diego, CA, USA, 2011, pp. 292–300.

[5] Y. Wang, S. Wang, and K. K. Lai, "A new fuzzy support vector machine to evaluate credit risk," *IEEE Trans. Fuzzy Syst.*, vol. 13, no. 6, pp. 820–831, Dec. 2005.

[6] S. C. Berger and F. Gleisner, "Emergence of financial intermediaries in electronic markets: The case of online P2P lending," *BuR Bus. Res.*, vol. 2, no. 1, pp. 39–65, May 2009.

[7] S. Ceyhan, X. Shi, and J. Leskovec, "Dynamics of bidding in a P2P lending service: Effects of herding and predicting loan success," in *Proc. ACM WWW*, Hyderabad, India, 2011, pp. 547–556.

[8] M. Herzenstein, U. M. Dholakia, and R. L. Andrews, "Strategic herding behavior in peer-to-peer loan auctions," *J. Interact. Mark.*, vol. 25, no. 1, pp. 27–36, Feb. 2011.

[9] M. Herzenstein, R. L. Andrews, U. Dholakia, and E. Lyandres, "The democratization of personal consumer loans? Determinants of success in online peer-to-peer lending communities," Working Paper. [Online]. Available: http://ssrn.com/abstract=1147856

[10] T. Stafinski, D. Menon, C. McCabe, and D. J. Philippon, "To fund or not to fund," *Pharmacoeconomics*, vol. 29, no. 9, pp. 771–780, Sep. 2011.

[11] M. Grinblatt, S. Titman, and R. Wermers, "Momentum investment strategies, portfolio performance, and herding: A study of mutual fund behavior," *Amer. Econ. Rev.*, vol. 85, no. 5, pp. 1088–1105, Dec. 1995.

[12] E. Lee and B. Lee, "Herding behavior in online P2P lending: An empirical investigation," *Electron. Commer. Res. Appl.*, vol. 11, no. 5, pp. 495–503, Sep. 2012.

[13] H. Markowitz, "Portfolio selection," *J. Finance*, vol. 7, no. 1, pp. 77–91, Mar. 1952.

[14] J. A. León and D. B. Tumpson, "Competition between two species for two complementary or substitutable resources," *J. Theor. Biol.*, vol. 50, no. 1, pp. 185–201, Mar. 1975.

[15] W. Nicholson and C. M. Snyder, *Microeconomic Theory: Basic Principles and Extensions*. Mason, OH, USA: Cengage Learn., 2011.

[16] E. Rosenberg and A. Gleit, "Quantitative methods in credit management: A survey," *Oper. Res.*, vol. 42, no. 4, pp. 589–613, Aug. 1994.

[17] H. Zhao, L. Wu, Q. Liu, Y. Ge, and E. Chen, "Investment recommendation in P2P lending: A portfolio perspective with risk management," in *Proc. IEEE ICDM*, Shenzhen, China, 2014, pp. 1109–1114.

[18] U. M. Fayyad and K. B. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in *Proc. IJCAI*, Chambéry, France, 1993, pp. 1022–1029.

[19] Q. Liu, Y. Ge, Z. Li, E. Chen, and H. Xiong, "Personalized travel package recommendation," in *Proc. IEEE ICDM*, Vancouver, BC, Canada, 2011, pp. 407–416.

[20] C. M. Turner, R. Startz, and C. R. Nelson, "A Markov model of heteroskedasticity, risk, and learning in the stock market," *J. Finance Econ.*, vol. 25, no. 1, pp. 3–22, Nov. 1989.

[21] Y. Cao, Y. Li, S. Coleman, A. Belatreche, and T. M. McGinnity, "Adaptive hidden Markov model with anomaly states for price manipulation detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 2, pp. 318–330, Feb. 2015.

[22] C. Zhu, H. Zhu, H. Xiong, P. Ding, and F. Xie, "Recruitment market trend analysis with sequential latent variable models," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, San Francisco, CA, USA, Aug. 2016, pp. 383–392.

[23] Y. Zhang, Y. Xiong, X. Kong, and Y. Zhu, "Netcycle: Collective evolution inference in heterogeneous information networks," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, San Francisco, CA, USA, 2016, pp. 1365–1374.

[24] S. Goldwater and T. Griffiths, "A fully Bayesian approach to unsupervised part-of-speech tagging," in *Proc. Annu. Meeting Assoc. Comput. Linguist.*, vol. 45. Prague, Czech Republic, 2007, pp. 744–751.

[25] B. Huai *et al.*, "Toward personalized context recognition for mobile users: A semisupervised Bayesian HMM approach," *ACM Trans. Knowl. Disc. Data*, vol. 9, no. 2, pp. 1–29, Nov. 2014.

[26] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 6, pp. 721–741, Nov. 1984.

[27] T. L. Griffiths and M. Steyvers, "Finding scientific topics," in *Proc. Nat. Acad. Sci.*, vol. 101. Washington, DC, USA, 2004, pp. 5228–5235.

[28] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inf. Theory*, vol. 13, no. 2, pp. 260–269, Apr. 1967.

[29] B. Kedem and K. Fokianos, *Regression Models for Time Series Analysis*, vol. 488. Hoboken, NJ, USA: Wiley, 2005.

[30] M. Deshpande and G. Karypis, "Item-based top-N recommendation algorithms," *ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 143–177, Jan. 2004.

[31] D. Maltz and K. Ehrlich, "Pointing the way: Active collaborative filtering," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, Denver, CO, USA, 1995, pp. 202–209.

[32] V. Kuppuswamy and B. L. Bayus, "Crowdfunding creative ideas: The dynamics of project backers in Kickstarter," Working Paper. [Online]. Available: http://ssrn.com/abstract=2234765

[33] A. Bachmann *et al.*, "Online peer-to-peer lending—A literature," *J. Internet Banking Commer.*, vol. 16, no. 2, pp. 1–18, 2011.

[34] L. Puro, J. E. Teich, H. Wallenius, and J. Wallenius, "Borrower decision aid for people-to-people lending," *Decis. Support Syst.*, vol. 49, no. 1, pp. 52–60, Apr. 2010.

[35] C.-T. Lu, S. Xie, X. Kong, and P. S. Yu, "Inferring the impacts of social media on crowdfunding," in *Proc. 7th ACM Int. Conf. Web Search Data Min.*, New York, NY, USA, 2014, pp. 573–582.

[36] S. Herrero-Lopez, "Social interactions in P2P lending," in *Proc. 3rd Workshop SNMA*, Paris, France, 2009, pp. 1–8.

[37] S. Freedman and G. Z. Jin, "Do social networks solve information problems for peer-to-peer lending? Evidence from prosper.com," Working Paper. [Online]. Available: http://hdl.handle.net/1721.1/77217

[38] X. D. Huang, Y. Ariki, and M. A. Jack, *Hidden Markov Models for Speech Recognition*, vol. 19. Edinburgh, U.K., Edinburgh Univ. Press, 1990.

[39] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[40] J. H. Albert and S. Chib, "Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts," *J. Bus. Econ. Stat.*, vol. 11, no. 1, pp. 1–15, Jan. 1993.

[41] J. D. Hamilton, "A new approach to the economic analysis of nonstationary time series and the business cycle," *Econometrica J. Econom. Soc.*, vol. 57, no. 2, pp. 357–384, Mar. 1989.

[42] D. R. Fredkin and J. A. Rice, "Bayesian restoration of single-channel patch clamp recordings," *Biometrics*, vol. 48, no. 2, pp. 427–448, Jun. 1992.

[43] B. G. Leroux and M. L. Puterman, "Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models," *Biometrics*, vol. 48, no. 2, pp. 545–558, Jun. 1992.

[44] Y. Li *et al.*, "Market demand oriented data-driven modeling for dynamic manufacturing system control," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 45, no. 1, pp. 109–121, Jan. 2015.

[45] X. Zhang *et al.*, "Human pose estimation and tracking via parsing a tree structure based human model," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 5, pp. 580–592, May 2014.

[46] H. Zhu, C. Liu, Y. Ge, H. Xiong, and E. Chen, "Popularity modeling for mobile Apps: A sequential approach," *IEEE Trans. Cybern.*, vol. 45, no. 7, pp. 1303–1314, Jul. 2015.

[47] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Stat.*, vol. 41, no. 1, pp. 164–171, Feb. 1970.

[48] J. Liechty, R. Pieters, and M. Wedel, "Global and local covert visual attention: Evidence from a Bayesian hidden Markov model," *Psychometrika*, vol. 68, no. 4, pp. 519–541, Dec. 2003.

[49] S. Guha, Y. Li, and D. Neuberg, "Bayesian hidden Markov modeling of array CGH data," *J. Amer. Stat. Assoc.*, vol. 103, no. 482, pp. 485–497, Jun. 2008.

[50] J. Gao and M. Johnson, "A comparison of Bayesian estimators for unsupervised hidden Markov model POS taggers," in *Proc. EMNLP*, Honolulu, HI, USA, 2008, pp. 344–352.

[51] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, no. 2, pp. 171–185, Apr. 1995.

Hongke Zhao (M'16) received the B.E. degree in software engineering from the South China University of Technology, Guangzhou, China, in 2013. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, University of Science and Technology of China, Hefei, China.

He was with the University of North Carolina at Charlotte, Charlotte, NC, USA, as a Research Scholar, for about one year. He is currently a Research Scholar with the Eller College of Management, University of Arizona, Tucson, AZ, USA. He has published several papers in refereed conference proceedings, such as ACM Conference on Knowledge Discovery and Data Mining, IEEE International Conference on Data Mining, and Springer International Conference on Database Systems for Advanced Applications. His current research interest includes data mining, with a focus on Internet finance such as P2P lending and Crowdfunding.

Qi Liu received the Ph.D. degree in computer science from the University of Science and Technology of China (USTC), Hefei, China.

He is an Associate Professor with USTC. He has published prolifically in refereed journals and conference proceedings, such as, the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS PART B: CYBERNETICS, *ACM Transactions on Information Systems*, *ACM Transactions on Knowledge Discovery from Data*, *ACM Transactions on Intelligent Systems and Technology*, ACM Conference on Knowledge Discovery and Data Mining, International Joint Conference on Artificial Intelligence, IEEE International Conference on Data Mining (ICDM), and ACM International Conference on Information and Knowledge Management. His current research interests include data mining and knowledge discovery.

Dr. Liu was a recipient of the ICDM-2011 Best Research Paper Award, the Special Prize of President Scholarship for Postgraduate Students, Chinese Academy of Sciences (CAS), and the Distinguished Doctoral Dissertation Award of CAS. He has served regularly in the program committees of a number of conferences, and is a Reviewer for the leading academic journals in his fields. He is a member of ACM.

Hengshu Zhu received the B.E. and Ph.D. degrees in computer science from the University of Science and Technology of China, Hefei, China, in 2009 and 2014, respectively.

He is currently a Senior Data Scientist with Big Data Laboratory, Baidu Research, Beijing, China. He has published prolifically in refereed journals and conference proceedings, including the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSCATIONS ON MOBILE COMPUTING, *ACM Transactions on Knowledge Discovery from Data*, ACM Conference on Knowledge Discovery and Data Mining, and the International Joint Conference on Artificial Intelligence. His current research interests include data mining and machine learning, with a focus on developing effective and efficient data analysis techniques for emerging big data intensive applications.

Dr. Zhu was a recipient of the Distinguished Dissertation Award, China Association for Artificial Intelligence in 2016, the Special Prize of President Scholarship for Postgraduate Students, Chinese Academy of Sciences in 2014, and the Best Student Paper Award of KSEM-2011 and WAIM-2013. He was regularly on the program committees of numerous conferences, and has served as a Reviewer for many top journals in relevant research fields. He is the member of ACM and the Communication Committee Member of CCF Task Force on Big Data.

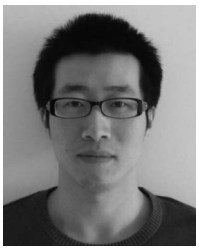
Yong Ge received the B.E. degree in information engineering from Xi'an Jiao Tong University, Xi'an, China, in 2005, the M.S. degree in signal and information processing from the University of Science and Technology of China, Hefei, China, in 2008, and the Ph.D. degree in information technology from Rutgers, the State University of New Jersey, Newark, NJ, USA, in 2013.

He is currently an Assistant Professor with the Eller College of Management, University of Arizona, Tucson, AZ, USA. His current research interests include data mining and business analytics. He has published prolifically in refereed journals and conference proceedings, such as the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, *ACM Transactions on Information Systems*, *ACM Transactions on Knowledge Discovery from Data*, *ACM Transactions on Intelligent Systems and Technology*, ACM Conference on Knowledge Discovery and Data Mining, SIAM International Conference on Data Mining, IEEE International Conference on Data Mining, and ACM RecSys.

Enhong Chen (SM'07) received the B.S. degree from Anhui University, Hefei, China, the M.S. degree from the Hefei University of Technology, Hefei, and the Ph.D. degree in computer science from the University of Science and Technology of China (USTC), Hefei.

He is currently a Professor and the Vice Dean of the School of Computer Science, the Vice Director of the National Engineering Laboratory for Speech and Language Information Processing, USTC. He has published lots of papers on refereed journals and conferences, including the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON MOBILE COMPUTING, ACM Conference on Knowledge Discovery and Data Mining (SIGKDD), IEEE International Conference on Data Mining (ICDM), Annual Conference on Neural Information Processing Systems, ACM International Conference on Information and Knowledge Management, and Nature *Communications*. His current research interests include data mining and machine learning, social network analysis, and recommender system.

Dr. Chen was a recipient of the National Science Fund for Distinguished Young Scholars of China, the Best Application Paper Award on SIGKDD-2008, and the Best Research Paper Award on ICDM-2011. He was on program committees of numerous conferences including ACM SIGKDD, IEEE ICDM, and SIAM International Conference on Data Mining.



Yan Zhu received the B.E. and M.E. degrees in computer science from the University of Science and Technology of China, Hefei, China, in 2013 and 2016, respectively.

He is currently a Quantitative Analyst with Laurion Capital Management, Shanghai, China. His research interests include statistical learning in risk management and securities investment, machine learning, and pattern recognition, with special applications for human behavior analysis.

Junping Du received the Ph.D. degree in computer science from the University of Science and Technology Beijing, Beijing, China.

She held a Post-Doctoral Fellowship with the Department of Computer Science, Tsinghua University, Beijing. She joined the School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, in 2006, where she is currently a Professor of computer science. She was a Visiting Professor with the Department of Computer Science, Aarhus University, Aarhus, Denmark, from 1996 to 1997. Her current research interests include artificial intelligence, data mining, intelligent management system development, and computer applications.

Dr. Du served as the Program Chair and the Program Co-chair for many international and domestic academic conferences, and has been a Vice General Secretary of Chinese Association for Artificial Intelligence, since 2004.