Frontiers of Information Technology & Electronic Engineering
www.zju.edu.cn/jzus; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

E-commerce business model mining and prediction*

Zhou-zhou HE[1], Zhong-fei ZHANG[‡1], Chun-ming CHEN[2], Zheng-gang WANG[2]

(*[1]Department of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China*)

(*[2]Alibaba Group, Hangzhou 310027, China*)

E-mail: zhouzhouhe@zju.edu.cn; zhongfei@zju.edu.cn; chunming.chencm@taobao.com; zhenggang.wangzg@alibabainc.com

Received May 5, 2015; Revision accepted Aug. 4, 2015; Crosschecked Aug. 7, 2015

Abstract: We study the problem of business model mining and prediction in the e-commerce context. Unlike most existing approaches where this is typically formulated as a regression problem or a time-series prediction problem, we take a different formulation to this problem by noting that these existing approaches fail to consider the potential relationships both among the consumers (consumer influence) and among the shops (competitions or collaborations). Taking this observation into consideration, we propose a new method for e-commerce business model mining and prediction, called EBMM, which combines regression with community analysis. The challenge is that the links in the network are typically not directly observed, which is addressed by applying information diffusion theory through the consumer-shop network. Extensive evaluations using Alibaba Group e-commerce data demonstrate the promise and superiority of EBMM to the state-of-the-art methods in terms of business model mining and prediction.

Key words: E-commerce, Business model prediction, Consumer influence, Social network, Sales prediction
doi:10.1631/FITEE.1500148 **Document code:** A **CLC number:** TP391

1 Introduction

Social behavior mining and prediction is concerned with discovering the relationship between typically a huge number of users and their corresponding social events along a timeline so that we can predict the future patterns of the social events. In this paper, we focus on a specific but very popular scenario of social behavior mining and prediction that has substantial impacts on business applications—business model mining and prediction from online sales data. With the rapid development of electronic commerce in recent years, there is a strong need and necessity to discover and predict the different business models for the large number of e-commerce shops in the Internet so that an e-commerce corporation that houses these shops (such as Alibaba Group) or the public investors may identify those shops with a good business model for future investment. Note that, in general, there are a number of variables to describe a business model for a business enterprise including the sales volume, revenues, and inventory consumption rate. However, from the perspective of the e-commerce corporation that houses these shops, there is only data for a part of the variables available as the data for the rest of the variables is not visible to the corporation. Consequently, in this paper, we study the special case of the business model mining and prediction problem for e-commerce shops—mining and predicting a shop's sales pattern. Note that we focus on the specific variable of sales of a shop as this is the only variable to describe the shop's business model that has the data available to the e-commerce corporation housing the shop. This is a non-trivial problem as accurately mining and predicting a shop's sales pattern depends upon not only the history sales data of the shop, but, more importantly, many other factors

‡ Corresponding author

* Project supported by the National Basic Research Program (973) of China (No. 2012CB316400), Zhejiang University - Alibaba Financial Joint Lab, Zhejiang Provincial Engineering Center on Media Data Cloud Processing and Analysis, China, and the US National Science Foundation (No. CCF-1017828)

ORCID: Zhou-zhou HE, http://orcid.org/0000-0003-3947-4011

such as the relationships between this shop and its peers (e.g., competition relationships) and relationships among the consumers of this shop. It is not as simple and trivial as a regression or time-series prediction from the shop's history data.

In the literature, the problem of mining and predicting business models for a business entity such as an e-commerce shop is typically solved through regression. However, all the existing approaches assume that the business entity in question is independent. In reality, however, this assumption is always violated as there always exist relationships between this business entity and its peers in the market (e.g., competition relationships). We argue that capturing these relationships among the business entities in the market (especially for those conducting similar business activities such as selling similar products) is essential in more accurately mining and predicting the business models. This is particularly true in the scenario of e-commerce where there are many shops competing each other in selling similar products or products with similar functions. The reason why the independence assumption is always used in the existing methods is that the relationships among the shops are typically not observed. Further, the relationships among the consumers are not observed either. These facts pose great challenges to the problem and add great complexity to the potential solutions to be developed to solve the problem with the consideration of these facts.

We also note that though information diffusion theory has been used in many areas, it is used only in the related literature to solve the problem in regression when the relationships are known among the entities (Bonchi *et al.*, 2011; Tsur and Rappoport, 2012). Here, since the relationships are not observed, this also poses a challenge to applying information diffusion theory to solve the problem.

Based on this observation, we propose a new method that combines consumer influence analysis and shop relationship mining with the regression using information diffusion theory as an effective solution to this problem. We call this new method EBMM. We highlight the contributions of this study as follows:

1. This is the first effort on e-commerce business model mining and prediction with the unobserved potential relationships among the consumers and among the shops in the market.

2. We develop an effective optimal solution, EBMM, to business model mining and prediction combining information diffusion theory with community analysis.

3. We demonstrate the effectiveness and promise of EBMM in the real-world business model prediction application through extensive evaluations in comparison with the state-of-the-art methods.

4. We show the effectiveness and promise of EBMM in the real-world consumer influence mining application.

2 Data modeling in e-commerce business

To show the promising performance of EBMM in the actual industrial applications, we identify a specific scenario in e-commerce business in collaboration with Alibaba Group, which is a leading e-commerce company in the world and provides us with real and huge online transaction data after anonymization.

2.1 Data collection

There are thousands of shops and millions of consumers as well as billions of commercial items traded everyday in the Internet through Alibaba Group with over 6 TB data generated everyday. To show the power of EBMM, we specifically select a particular type of commercial item, women's fashion clothes, for this study. Moreover, to conduct an extensive evaluation, we randomly select a small portion of the sales data generated within one year (2012), involving over 16 millions of consumers.

In the business operation for the shops housed by Alibaba, it is typical that one shop sells only one type of commercial item. This is particularly true for the data we have collected for women's clothes sales where there is a one-to-one mapping between a specific shop and the corresponding specific fashion type of women's clothes. To further study the business activities of those significant shops, we focus on the five most popular commercial item categories of woman fashion clothes in 2012, i.e., one-piece dress, T-shirt, chiffon shirt, blouse, and sleeveless shirt, with their sales volumes and amounts for 20 continuous weeks in the timeline of 2012, and define those significant shops as those with their total sales volume of the clothes over 1000 pieces or higher.

In addition, we collect all the consumers who have bought the clothes from these shops. Fig. 1a describes the collected data represented as the general relational data (Long *et al.*, 2007). The resulting five relational datasets corresponding to the five categories of women's clothes sold in the time domain are used in our evaluation.

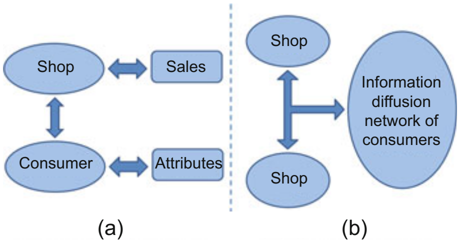


Fig. 1 Data model for the e-commerce business: (a) original data model; (b) EBMM data model

2.2 Data processing

Before we develop the EBMM framework, it is important to preprocess the data to fully capture the influence relationships among the shops, among the consumers, and between the shops and consumers. The preprocessing is conducted to capture two specific observations from the data.

First, note that there exist relationships among the shops, in particular, the competition and/or the collaboration relationships for the shops that sell clothes of the same category. To simplify the capturing of this type of relationship, for each shop we identify its 10 most similar shops from all its peer shops, where the similarity is defined through key word matching from the names of the clothes sold by the shops. We call these similar shops the 'related' shops. Second, given the huge number of consumers, it is extremely expensive to compute their potentially possible pairwise relationships. Further, this would make the information diffusion network very sparse. On the other hand, individual consumers may have their individual shopping behaviors, resulting in a strong noise in the information diffusion network. To focus on the typically representative shopping behavior of the consumers, we pick up the 100 000 consumers who have the strongest spending powers and apply k-means to their attribute space to generate 500 consumer clusters as the consumer representatives. In the rest of the paper, the word 'consumer' refers to a consumer cluster and the attributes of a consumer are the corresponding attributes of the center of the consumer cluster. Consequently, the information diffusion network is developed using these 500 consumers, with the simplified data representation shown in Fig. 1b.

3 Framework

In this section, we first make a problem statement to give a clear definition about what exactly the problem is. We then propose the EBMM framework combining consumer relationship analysis and shop relationship analysis with information diffusion theory, which is achieved as an iterative optimization method.

We begin with defining the notations used in this paper. Lowercase letters are used to represent the scalars. Capital boldface letters are used to represent the matrices and lowercase boldface letters the vectors. Denote $\{\boldsymbol{x}_m\}$ as the set of vectors with $\boldsymbol{x}_m$ representing the mth element in the set.

3.1 Problem statement

Assume that we are given a collection of data about N different shops and M different consumers. The unit in the timeline is day. The sales of the nth shop in T consecutive days are represented as a set of vectors $\{\boldsymbol{y}_n\}_{n=1}^{N}$, $\boldsymbol{y}_n = (y_{n,1}, y_{n,2}, \ldots, y_{n,T})^{\mathrm{T}} \in \mathbb{R}^{T\times 1}$. In addition, we collect the time information of transactions for the shops. For example, $\tau(i,n)$ denotes the time at which the ith consumer makes transactions with the nth shop. To represent the relationships among the shops, for the nth shop, we collect the set of its related shops as $\mathrm{Nb}(n)$.

After the collection of the transaction data between the shops and the consumers, we have the following two features to describe the shops:

Definition 1 (Internal influence factor) The internal influence factor is the attribute to describe the consumers influence in a shop. We denote the internal influence factor of the nth shop on the tth day as $\boldsymbol{a}_{n,t} \in \mathbb{R}^{M^2\times 1}$. Specifically, for the ith and jth consumers, if they make transactions with the nth shop in between the $(t-t_{\mathrm{r}})$th and tth days while the time interval of the transactions between the ith and jth consumers is no more than Δt days, we set the $((i-1)M+j)$th element of $\boldsymbol{a}_{n,t}$ to 1. Otherwise, we set this element to 0.

Definition 2 (External influence factor) The external influence factor is the attribute to describe the consumer influence among the shops. We denote the external influence factor of the nth shop on the tth day as $\boldsymbol{b}_{n,t} \in \mathbb{R}^{M^2\times 1}$. Specifically, for the ith and jth consumers, if the jth consumer makes transactions with the nth shop and the ith consumer makes transactions with the n'th shop (the related shop of the nth shop) in between the $(t-t_{\mathrm{r}})$th and tth days, while the time interval of the transactions between the ith and jth consumers is no more than Δt days, set the $((i-1)M+j)$th element of $\boldsymbol{b}_{n,t}$ to 1. Otherwise, set this element to 0.

The detailed formulations of the internal influence factor and external influence factor are described as follows:

$$\begin{aligned}&\boldsymbol{a}_{n,t}((i-1)M+j)\\&=\begin{cases}1, & \tau(i,n),\tau(j,n)\in[t-t_{\mathrm{r}},t],\\ & \tau(j,n)-\tau(i,n)\in(0,\Delta t],\\ 0, & \text{otherwise},\end{cases}\end{aligned}\tag{1}$$

$$\begin{aligned}&\boldsymbol{b}_{n,t}((i-1)M+j)\\&=\begin{cases}1, & \tau(i,n'),\tau(j,n)\in[t-t_{\mathrm{r}},t],\\ & \tau(j,n)-\tau(i,n')\in(0,\Delta t],\\ 0, & \text{otherwise}.\end{cases}\end{aligned}\tag{2}$$

Unlike the traditional approaches using the static features (e.g., the production's titles or prices) to describe the shops, the above two features can be used to capture the shops' dynamic sales behavior in the whole time domain under the influence of the consumers.

Consequently, the problem we investigate in this study is that given a dataset consisting of N shops and their features $\{\boldsymbol{a}_{n,t}\}$ and $\{\boldsymbol{b}_{n,t}\}$, predict the shops' business model and discover the latent influence model among the consumers. Table 1 summarizes the definitions of the symbols used in this paper.

3.2 Overall approach

To predict the shops' business model, it is important to establish a mapping from the shops' features to the shops' sales. However, there are two issues to consider in the real transaction data: (1) The sales volume of a shop may typically experience an unpredictable behavior with a large fluctuation for a certain period of time and a smooth volume for another period of time; (2) Due to the typical competition and cooperation relationships among the shops, the independence assumption is invalid (Myers and Leskovec, 2012).

Table 1 Detailed description of the variables

Variable	Description
N	Number of shops
M	Number of consumers
T	Number of days
$y_{n,t}$	Sales of the nth shop in the tth day
$\tau(i,n)$	Time of the transactions
$\mathrm{Nb}(n)$	The set of the related shops
t_{r}	Time interval
Δt	Maximum influence duration
$\boldsymbol{a}_{n,t}$	Internal influence factor
$\boldsymbol{b}_{n,t}$	External influence factor
$x_{n,t}$	True part of the sales
$z_{n,t}$	Noise part of the sales
$u_{n,t}$	Auxiliary variable
K	Number of clusters in the shops
$p_{n,k}$	Probability of the shop belonging to the cluster
$\boldsymbol{w}_k$	Internal regression vector
$\boldsymbol{v}_k$	External regression vector
λ_1	Constraint weight of the true part
λ_2	Constraint weight of the noise part
ρ	Penalty parameter

To address the first issue, we split the sales volume of a shop into two parts, true part and noise part, with the true part exhibiting a smooth development according to a certain distribution over the time and the noise part collecting all the unpredictable fluctuations over the time. Consequently, we take separate regressions to these two parts with different constraints, as opposed to taking the whole sales volume as one regression. To address the second issue, we assume that there exists the clustering structure in the shops and use the mixtures of the linear regression model to predict the true part and noise part of the sales, separately. With these two techniques, an overall solution is developed to solve the following target function:

$$\begin{aligned}\min\ &\sum_{n=1}^{N}\left\{\sum_{t=1}^{T}\mathrm{loss}(x_{n,t},\boldsymbol{a}_{n,t},\{\boldsymbol{w}_k\})\right.\\&+\sum_{t=1}^{T}\mathrm{loss}(z_{n,t},\boldsymbol{b}_{n,t},\{\boldsymbol{v}_k\})\\&\left.+\lambda_1\sum_{t=2}^{T}|x_{n,t}-x_{n,t-1}|+\lambda_2\sum_{t=1}^{T}|z_{n,t}|\right\}\\ \text{s.t.}\ & y_{n,t}=x_{n,t}+z_{n,t},\end{aligned}\tag{3}$$

where $x_{n,t}$ and $z_{n,t}$ denote the true part and noise part of the nth shop in the tth day, respectively, and $\{\boldsymbol{w}_k\}$ and $\{\boldsymbol{v}_k\}$ denote the sets of internal and external regression vectors, respectively, with the detailed definitions given in Eq. (4). The item $\sum_{t=2}^{T}|x_{n,t}-x_{n,t-1}|$ is the constraint for smoothing the true part of the nth shop's sales. It is used in fused lasso for regularization (Tibshirani *et al.*, 2005; Friedman *et al.*, 2007). The loss function loss$(\cdot)$, represented as the mixtures of the linear regression, establishes the mapping from the shops' features to the shops' sales. The specific definition of this function is described as follows:

$$\text{loss}(x,\boldsymbol{c},\{\boldsymbol{h}_k\})=\frac{1}{2}\left(x-\sum_{k=1}^{K}p_{x,k}\boldsymbol{h}_k^{\mathrm{T}}\boldsymbol{c}\right)^2$$
$$\text{s.t.}\ \sum_{k=1}^{K}p_{x,k}=1,\ p_{x,k}>0, \tag{4}$$

where $\boldsymbol{c}$ and x denote the predictor variables (features) and the response variables, respectively, $p_{x,k}$ denotes the probability of the sample belonging to the kth cluster, and $\boldsymbol{h}_k$ denotes the regression vector of the kth cluster. Fig. 2 shows an exemplary scenario for the process of making the prediction.

3.3 Algorithm development

The specific algorithm development of the EBMM framework is based on the considerations of incorporating the strategies to address the two issues in the literature through iteratively alternating two steps, sales splitting and regression learning.

3.3.1 Sales splitting

Given variables $\{\boldsymbol{w}_k\}$, $\{\boldsymbol{v}_k\}$, and $\{p_{n,k}\}$, we aim to split the sales in the nth shop in the tth day, $y_{n,t}$, into true part $x_{n,t}$ and noise part $z_{n,t}$. Now the target function in Eq. (3) is represented as

$$\min\ f(\{x_{n,t}\})+g(\{z_{n,t}\})$$
$$\text{s.t.}\ y_{n,t}=x_{n,t}+z_{n,t}, \tag{5}$$

where $f(\cdot)$ and $g(\cdot)$ are the functions to constrain $\{x_{n,t}\}$ and $\{z_{n,t}\}$, respectively.

The target function in Eq. (5), described as the combination of one fused lasso and one common lasso subject to the linear constraints, is difficult to solve in general. Therefore, we solve the problem using an ADMM (alternating directions method of multipliers) algorithm (Boyd *et al.*, 2011). Eq. (5) can be rewritten as the form of a scaled augmented Lagrangian:

$$L_\rho=f(\{x_{n,t}\})+g(\{z_{n,t}\})+\frac{\rho}{2}\sum_{n,t}(x_{n,t}+z_{n,t}-y_{n,t}+u_{n,t})^2, \tag{6}$$

where $\{u_{n,t}\}$ are the auxiliary variables and ρ is the penalty parameter. Specifically, the optimal solution corresponding to Eq. (6) is achieved by iterating

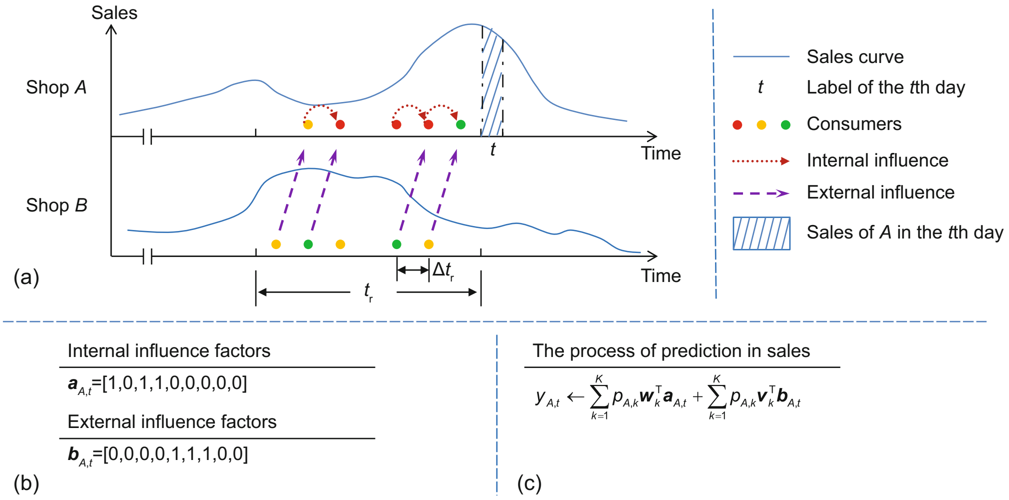

Sales
Shop A
Shop B
t
Time
Time
Δt_r
t_r
(a)
Sales curve
t Label of the tth day
● ● ● Consumers
Internal influence
External influence
▨ Sales of A in the tth day
Internal influence factors
a_{A,t}=[1,0,1,1,0,0,0,0,0]
External influence factors
b_{A,t}=[0,0,0,0,1,1,1,0,0]
(b)
The process of prediction in sales
$y_{A,t}\leftarrow\sum_{k=1}^{K}p_{A,k}\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{a}_{A,t}+\sum_{k=1}^{K}p_{A,k}\boldsymbol{v}_k^{\mathrm{T}}\boldsymbol{b}_{A,t}$
(c)

Fig. 2 An example of predicting the business model: (a) illustration of the business process in shops *A* and *B*; (b) the internal influence factor and external influence factor of shop *A* on the *t*th day; (c) formulation of prediction in shop *A*'s sales

three steps. At the lth iteration, they are as follows:

$$\{x_{n,t}\}^{l+1} \leftarrow \arg\min_{x} \Bigg\{ f(\{x_{n,t}\}) + \frac{\rho}{2}\sum_{n,t}\left(x_{n,t}+z_{n,t}^{l}-y_{n,t}+u_{n,t}^{l}\right)^2 \Bigg\}, \quad (7)$$

$$\{z_{n,t}\}^{l+1} \leftarrow \arg\min_{z} \Bigg\{ g(\{z_{n,t}\}) + \frac{\rho}{2}\sum_{n,t}\left(x_{n,t}^{l+1}+z_{n,t}-y_{n,t}+u_{n,t}^{l}\right)^2 \Bigg\}, \quad (8)$$

$$\{u_{n,t}\}^{l+1} \leftarrow \{u_{n,t}\}^{l} + (\{x_{n,t}\}^{l+1}+\{z_{n,t}\}^{l+1}-\{y_{n,t}\}). \quad (9)$$

We now give the detailed description of the ADMM algorithm:

1. Initialize the variables: $x_{n,t}=y_{n,t}$, $z_{n,t}=0$, $u_{n,t}=0$ for $l=0$.

2. Select the parameters: λ_1, λ_2, and ρ.

3. For $l=1,2,\ldots$ until convergence:

3.1. For the nth shop, update $\{x_{n,t}\}^{l+1}$ as the minimizer of

$$\sum_{t=1}^{T}\mathrm{loss}_{\boldsymbol{w}}(x_{n,t},\boldsymbol{a}_{n,t})+\lambda_1\sum_{t=2}^{T}|x_{n,t}-x_{n,t-1}| + \frac{\rho}{2}\sum_{n,t}\left(x_{n,t}+z_{n,t}^{l}-y_{n,t}+u_{n,t}^{l}\right)^2. \quad (10)$$

Further, we simplify the above equation as

$$\min\ \frac{1}{2}\sum_{t=1}^{T}(x_{n,t}-\hat{x}_{n,t})^2+\frac{\lambda_1}{\rho+1}\sum_{t=2}^{T}|x_{n,t}-x_{n,t-1}|$$
$$\text{s.t.}\quad \hat{x}_{n,t}=\frac{1}{\rho+1}\sum_{k=1}^{K}p_{n,k}\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{a}_{n,t} + \frac{\rho}{\rho+1}\left(y_{n,t}-z_{n,t}^{l}-u_{n,t}^{l}\right). \quad (11)$$

In Eq. (11) the optimal target is to solve the problem of fused lasso. Therefore, we obtain the solution to $\{x_{n,t}\}^{l+1}$ by using the path algorithm (Hoefling, 2010).

3.2. For the nth shop, update $\{z_{n,t}\}^{l+1}$ as the minimizer of

$$\sum_{t=1}^{T}\mathrm{loss}_{\boldsymbol{v}}(z_{n,t},\boldsymbol{b}_{n,t})+\lambda_2\sum_{t=1}^{T}|z_{n,t}| + \frac{\rho}{2}\sum_{n,t}\left(x_{n,t}^{l+1}+z_{n,t}-y_{n,t}+u_{n,t}^{l}\right)^2. \quad (12)$$

Further, we simplify the above equation as

$$\min\ \frac{1}{2}\sum_{t=1}^{T}(z_{n,t}-\hat{z}_{n,t})^2+\frac{\lambda_2}{\rho+1}\sum_{t=1}^{T}|z_{n,t}|$$
$$\text{s.t.}\quad \hat{z}_{n,t}=\frac{1}{\rho+1}\sum_{k=1}^{K}p_{n,k}\boldsymbol{v}_k^{\mathrm{T}}\boldsymbol{b}_{n,t} + \frac{\rho}{\rho+1}\left(y_{n,t}-x_{n,t}^{l+1}-u_{n,t}^{l}\right). \quad (13)$$

In Eq. (13) the optimal target is to solve the problem of lasso. Therefore, we obtain the analytical solution to $\{z_{n,t}\}^{l+1}$ by using the soft-thresholded function (Donoho and Johnstone, 1995):

$$z_{n,t}^{l+1}=\mathrm{sign}(\hat{z}_{n,t})\left(|\hat{z}_{n,t}|-\frac{\lambda_2}{\rho+1}\right)_{+}. \quad (14)$$

3.3. Update $u_{n,t}^{l+1}$ as $u_{n,t}^{l}+\rho\left(x_{n,t}^{l+1}+z_{n,t}^{l+1}-y_{n,t}\right)$ for the nth shop in the tth day.

3.3.2 Regression learning

Given $\{x_{n,t}\}$ and $\{z_{n,t}\}$, we aim to discover the shops' membership $\{p_{n,k}\}$ and build the mappings $\{\boldsymbol{w}_k\}$ and $\{\boldsymbol{v}_k\}$. Now the target function in Eq. (3) is represented as follows:

$$\min\sum_{n=1}^{N}\sum_{t=1}^{T}\mathrm{loss}_{\boldsymbol{w}}(x_{n,t},\boldsymbol{a}_{n,t})+\sum_{n=1}^{N}\sum_{t=1}^{T}\mathrm{loss}_{\boldsymbol{v}}(z_{n,t},\boldsymbol{b}_{n,t}). \quad (15)$$

Considering the target function in Eq. (15), the optimal solution is achieved by iterating the following two steps:

1. Updating $\{\boldsymbol{w}_k\}$ and $\{\boldsymbol{v}_k\}$, given $\{p_{n,k}\}$

Since Eq. (15) is the combination of two mixtures of linear regression, we obtain the optimal solutions to $\{\boldsymbol{w}_k\}$ and $\{\boldsymbol{v}_k\}$ by solving the following two linear equations:

$$\left(\sum_{t=1}^{T}\boldsymbol{A}_t^{\mathrm{T}}\boldsymbol{P}_k^{\mathrm{T}}\boldsymbol{P}_k\boldsymbol{A}_t\right)\boldsymbol{w}_k=\sum_{t=1}^{T}\boldsymbol{A}_t^{\mathrm{T}}\boldsymbol{P}_k^{\mathrm{T}}\boldsymbol{P}_k\boldsymbol{x}_t, \quad (16)$$

$$\left(\sum_{t=1}^{T}\boldsymbol{B}_t^{\mathrm{T}}\boldsymbol{P}_k^{\mathrm{T}}\boldsymbol{P}_k\boldsymbol{B}_t\right)\boldsymbol{v}_k=\sum_{t=1}^{T}\boldsymbol{B}_t^{\mathrm{T}}\boldsymbol{P}_k^{\mathrm{T}}\boldsymbol{P}_k\boldsymbol{z}_t, \quad (17)$$

where $\boldsymbol{A}_t,\boldsymbol{B}_t\in\mathbb{R}^{N\times M^2}$ with the nth row of $\boldsymbol{A}_t$ and $\boldsymbol{B}_t$ being $\boldsymbol{a}_{n,t}^{\mathrm{T}}$ and $\boldsymbol{b}_{n,t}^{\mathrm{T}}$, respectively, $\boldsymbol{x}_t,\boldsymbol{z}_t\in\mathbb{R}^{N\times 1}$ with the nth element of $\boldsymbol{x}_t$ and $\boldsymbol{z}_t$ being $x_{n,t}$ and $z_{n,t}$, respectively, and $\boldsymbol{P}_k\in\mathbb{R}^{N\times N}$ is the diagonal matrix with the element of $\boldsymbol{P}_k$ at the nth row and nth column being $p_{n,k}$.

2. Updating $\{p_{n,k}\}$ given $\{\boldsymbol{w}_k\}$ and $\{\boldsymbol{v}_k\}$

Since using the traditional linear programming method to compute $\boldsymbol{P}$ is complex, we obtain the optimal solution to $\{p_{n,k}\}$ by solving the following simple equation:

$$p_{n,k} = \frac{\sum_{t=1}^{T} \exp\left(-\epsilon_{k,n,t}^2\right)}{\sum_{k=1}^{K}\sum_{t=1}^{T} \exp\left(-\epsilon_{k,n,t}^2\right)}, \tag{18}$$

where $\epsilon_{k,n,t} = \boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{a}_{n,t} + \boldsymbol{v}_k^{\mathrm{T}}\boldsymbol{b}_{n,t} - y_{n,t}$.

The shops' memberships in the testing set are obtained through their related shops, and the shops' sales are predicted by the following equation:

$$\widetilde{y}_{n,t} = \sum_{k=1}^{K} p_{n,k}\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{a}_{n,t} + \sum_{k=1}^{K} p_{n,k}\boldsymbol{v}_k^{\mathrm{T}}\boldsymbol{b}_{n,t} \tag{19}$$

$$\text{s.t.}\ \ p_{n,k} = \frac{1}{|n'|}\sum_{n\in n'} p_{n',k}, \quad n' = \mathrm{Nb}(n)\cap \mathrm{TrainingSet}.$$

Now the EBMM algorithm is described in detail in Algorithm 1.

Algorithm 1 EBMM

```
Input: the shops' features {$\boldsymbol{a}_{n,t}$}, {$\boldsymbol{b}_{n,t}$} and the shops' sales {$y_{n,t}$}.
Output: the shops' true part {$x_{n,t}$} and noise part {$z_{n,t}$}, the regression vectors {$\boldsymbol{w}_k$} and {$\boldsymbol{v}_k$}, and the shops' memberships {$p_{n,k}$}.
loop
    Sales splitting:
    Use the ADMM algorithm to obtain the optimal solution of {$x_{n,t}$} and {$z_{n,t}$}
    Regression learning:
    loop
        Use Eqs. (16) and (17) to obtain the optimal solution of {$\boldsymbol{w}_k$} and {$\boldsymbol{v}_k$}, respectively
        Use Eq. (18) to obtain the optimal solution of {$p_{n,k}$}
    end loop
end loop
Predict the shops' sales in the testing set
```

4 Experiments

In this section, we report the extensive evaluations using the datasets in the real e-commerce applications discussed in Section 2. The experiments are composed of two parts: business model prediction and consumer influence investigation.

4.1 Business model prediction

We first give a brief introduction to the competing methods used in the business model prediction evaluations. The competing models studied in this work are:

1. Traditional regression learning (TRL): For the nth shop in the tth day, this approach seeks a uniform mapping to predict the sales $y_{n,t}$ based on features $\boldsymbol{a}_{n,t}$ and $\boldsymbol{b}_{n,t}$.

2. Time series analysis (TSA) (Box, 2008): For the nth shop in the tth day, this approach seeks a uniform mapping to predict the sales $y_{n,t}$ based on historical sales $\{y_{n',t'}\}, n' \in n \cup \mathrm{Nb}(n), t' \in [t-t_{\mathrm{r}}, t]$.

3. Linear influence model (LIM) (Yang and Leskovec, 2010): For the nth shop in the tth day, this approach seeks a uniform mapping to predict sales $y_{n,t}$ based on the influence functions of the related consumers within a certain time interval.

In this work, we consider the first two models as the baseline ones. In addition, we give the definition of accuracy for the evaluations of the prediction quality of an approach in the experiments:

$$\text{accuracy} = \frac{1}{N_{\text{test}}T_{\text{test}}}\sum_{n=1}\sum_{t=1}\mathrm{Score}\left(\widetilde{y}_{n,t}, y_{n,t}\right),$$

$$\mathrm{Score}\left(\widetilde{y}_{n,t}, y_{n,t}\right) = \begin{cases} 1, & \dfrac{|\widetilde{y}_{n,t} - y_{n,t}|}{y_{n,t}} \leq \text{thd}, \\ 0, & \text{otherwise}, \end{cases}$$

where N_{test} and T_{test} are the numbers of shops and days in the testing set, respectively, and thd is set to 0.2 in the experiments. Further, to facilitate a systematic evaluation, we partition a dataset into different training sets and testing sets with different testing ratios. For example, the 20% testing ratio means that 20% of the samples of the dataset are used as the testing dataset and the remaining samples are used as the training set. We have tried different parameter values and found that EBMM is not sensitive to these different parameter values. In the experiments, we set $t_{\mathrm{r}} = 7$, $\Delta t = 2$, $\lambda_1 = 50$, $\lambda_2 = 100$, $\rho = 1$, $K = 4$, $M = 500$, and $N = 10\ 024$.

To compare EBMM and the competing models, we predict these models for the five datasets described in Section 2. Fig. 3 shows the accuracy of using the four models for these five datasets with different testing ratios. The accuracy of EBMM is about 0.81 on average and the performance of EBMM is about 14.6% better than that of TSA on average and about 25.2% better than that of LIM on average for the five datasets. In particular, the accuracy of EBMM is much higher than those of the others in the dataset of blouse. The accuracy of TRL is always the lowest with a large margin to even the next best method. This indicates that traditional regression completely fails in accurately predicting a business model. The second best method is the time-series prediction method. In this evaluation, it is also clear that the time-series approaches do not work as they fail to consider the consumer relationship information and the shops' relationship information, which turns out essential for giving an accurate prediction for an e-commerce business model according to this evaluation. Further, when the testing ratio increases from 10% to 60%, the performances of all the methods approximately remain stable for all the five datasets. However, when the testing ratio increases from 70% to 90%, EBMM has greater stability than the other methods. This indicates that EBMM can use a much smaller amount of sample data to achieve a more accurate prediction for a large-scale dataset.

While Fig. 3 demonstrates the overall systematic evaluations for EBMM's superiority to the existing methods, we further show EBMM's superiority with five prediction examples from five individual shops' sales randomly selected from the dataset in Fig. 4, where we use the first 8 weeks' data as the history and the next 12 weeks' data as the ground truth. For comparison, predictions are made using three models—since TRL always achieves the worst performance in the systematic evaluation, we drop it here. Clearly, EBMM's prediction is very close to the ground truth, while LIM and TSA fail in making reasonable predictions, further indicating that simple regressions or conventional time-series predictions do not work at all.

To further show the robustness and superiority of EBMM and justify the rationale why we split the sales volume into two parts and take separate regressions, we report the prediction accuracy of these four methods in Fig. 5 for a period of time during which large fluctuations have occurred. Although these four methods have overall relatively poor performances compared with those for the whole timeline reported in Fig. 3, the accuracy of EBMM is still about 80% better than that of TSA on average and about 90% better than that of LIM on average, with TRL completely failing. This evaluation shows that EBMM has strong robustness even for data with a large unpredictability. The noise part of the shops' sales $\{z_{n,t}\}$ reflects the sales' fluctuation very well.

4.2 Consumer shopping influence investigation

While e-commerce business model prediction is an important goal of this study, knowledge discovery from the e-commerce data is also an important

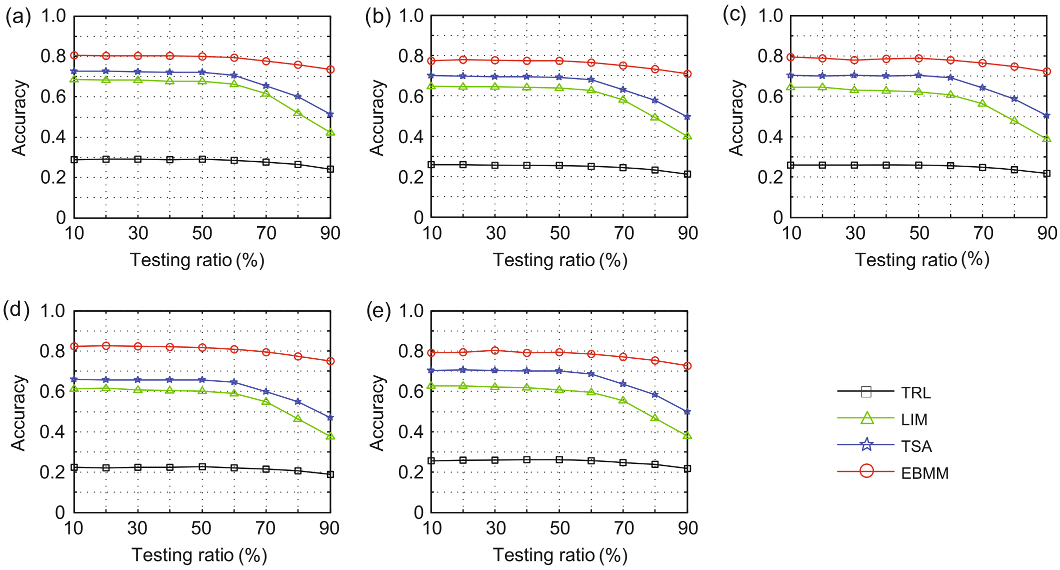


Fig. 3 Accuracy of sales prediction using four models for five different datasets: (a) one-piece dress; (b) T-shirt; (c) chiffon shirt; (d) blouse; (e) sleeveless shirt

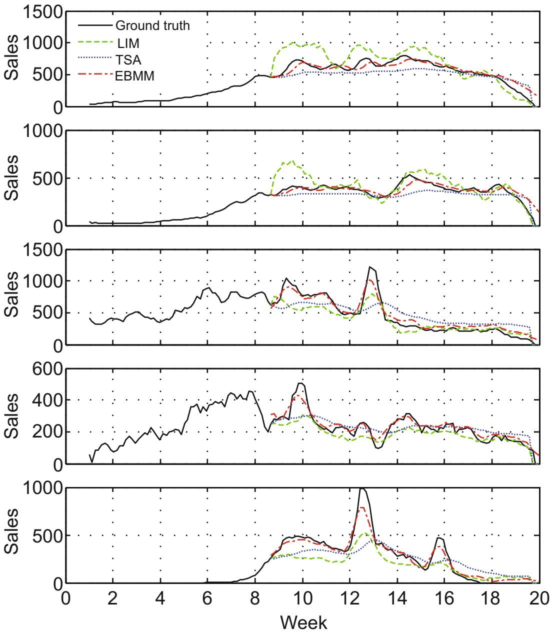


Fig. 4 Examples of the individual sales prediction for five different shops

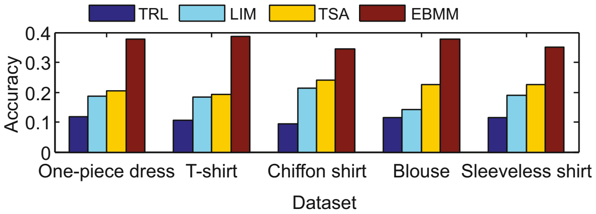


Fig. 5 Accuracy of sales prediction using four models for a period of time during which large fluctuations occur

goal for business model mining. Specifically, we are interested in influence propagation for the consumer shopping patterns in the consumer networks. To achieve this goal, we first partition the consumers into K_{c} clusters through k-means as it is considered that there are common shopping patterns for groups of the consumers. We call each such consumer cluster a community with the set of consumers in the ith community denoted as $C_i, i = 1, 2, \ldots, K_{\mathrm{c}}$. To appropriately measure the shopping influence of one consumer community to another, we explicitly define the metric $\mathrm{InfluencePower}_{i,j}^{k}$ as the influence power from consumer community i to consumer community j at the kth regression:

$$\mathrm{InfluencePower}_{i,j}^{k} = \frac{1}{|C_i| \cdot |C_j|} \sum w_{m,m'}^{k}$$
$$\text{s.t.} \quad m \in C_i, m' \in C_j,$$

where $|C_i|$ denotes the number of consumers in the ith community and $w_{m,m'}^{k}$ the $((m-1)M + m')$th element of $\boldsymbol{w}_k$.

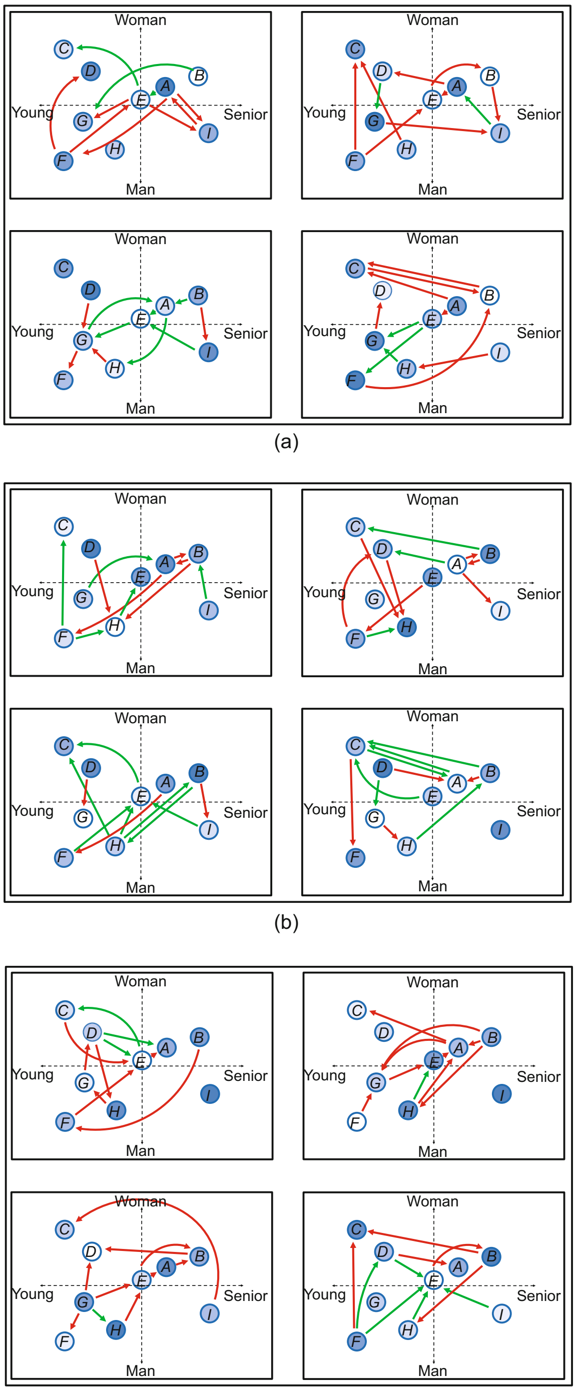
Since we have no prior knowledge about K_{c}, the number of consumer communities, we have exhaustively tried all the different values of K_{c} and have found that when $K_{\mathrm{c}} = 9$ the resulting communities have the best physical interpretations for the consumer shopping behavior. Specifically, we define the different and specific consumer communities in terms of their specific shopping preferences (Table 2). Furthermore, to visualize the consumer communities and the influence among them in a two-dimensional (2D) space, we collect data with two pieces of statistical information (i.e., the young/senior ratio and man/woman ratio in the communities) to help visualize the mining results in this 2D space. In the rest of this experiment, we always set $K_{\mathrm{c}} = 9$.

To graphically best describe the mined influence power among the consumer communities under different circumstances, we map and embed the mined influence power data into a 2D space with the first dimension being the attribute 'age' and the second dimension the attribute 'sex'. While the age attribute is a continuous variable running from young age to senior age with a specific value indicating the average age of all the consumers in a community, the sex attribute is also a continuous variable running from man (with a coordinate value 0) to woman (with a coordinate value 1) with a specific value indicating the ratio of the number of female consumers to the total number of consumers in a community (e.g., the middle point in this dimension means that the ratio of female consumers to male consumers in a community is 1:1).

Table 2 The detailed description of the consumer communities

Community	Consumer preferences
A	Snack, cosmetics, women's shoes
B	Children's clothes, women's shoes, men's shoes, men's clothes
C	Online game, virtual community, young woman
D	All the commodities except luxury
E	Children's clothes, cosmetics, car services, phone, sport equipment
F	Online game, virtual community, young man
G	Milk powder, travel services, cosmetics, phone
H	Children's clothes, milk powder, travel services, cosmetics, phone
I	Travel services, car services, jewelry

Figs. 6 and 7 show the detailed and specific mining results for the influence power propagation from one consumer community to another in the five datasets.

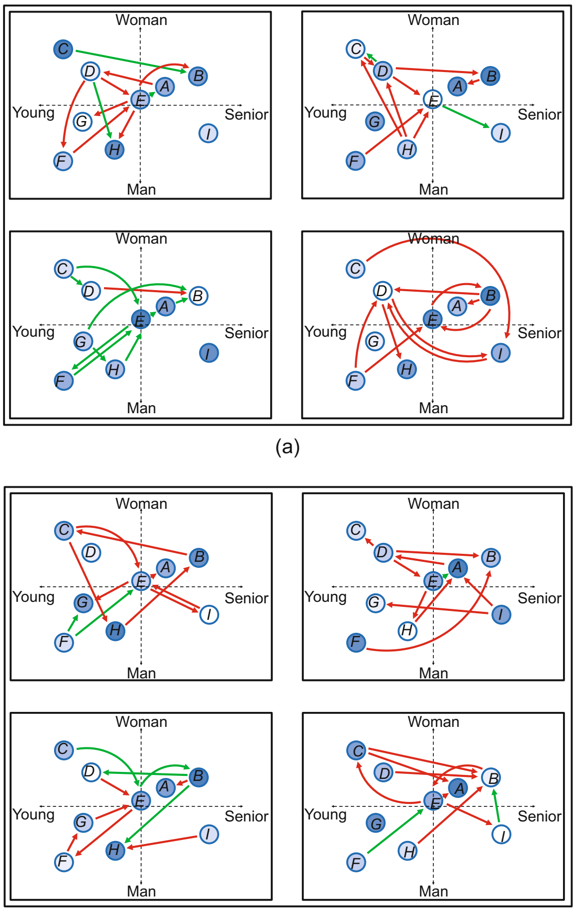


(a)

(b)

(c)

Fig. 6 Influence power mining among the consumer communities in one-piece dress (a), T-shirt (b), and chiffon shirt (c) datasets. References to color refer to the online version of this figure



(a)

(b)

Fig. 7 Influence power mining among the consumer communities in blouse (a) and sleeveless shirt (b) datasets. References to color refer to the online version of this figure

Here a circled node denotes a specific consumer community with the symbol inside the node corresponding to the community defined in Table 2. The red arrow means a positive influence; e.g., a red arrow from consumer community A to consumer community I means that if consumers in community A have bought the merchandise of the corresponding dataset at this time unit, it is likely that consumers in community I will buy the same merchandise at the next time unit. The green arrow means a negative influence; e.g., a green arrow from consumer community B to consumer community G means that if consumers in community B have bought the merchandise of the corresponding dataset at this time unit, it is likely that consumers in community G will not buy the same merchandise at the next time unit. Since the consumer shopping influence of one consumer community on another can also include the scenario of self-influence within a consumer community, the shade of a circle in the figures represents the strength of the community's self-influence with a darker shade for a stronger self-influence strength.

From the mining results in Figs. 6 and 7, we have the following observations:

1. The nine consumer communities essentially are almost evenly represented by different consumers with different age groups and different sex groups.

2. Different consumer shopping influence patterns (including different self-influence patterns) exist for different merchandise items (e.g., overall consumers have more negative consumer shopping influence on each other than positive consumer shopping influence on each other for T-shirts but have more positive consumer shopping influence on each other than negative consumer shopping influence on each other for chiffon shirts).

3. More consumer shopping influence (either positive or negative) exists within the same age group or within the same sex group than across different age groups or across different sex groups.

Since, to our best knowledge, this is the first effort for this type of consumer influence mining, we do not have a comparison evaluation with the literature.

5 Related work

In this section, two areas most related to this study are reviewed.

5.1 Modeling information diffusion

In recent years, many approaches about predicting and mining the behaviors of users have been proposed based on modeling information diffusion. Most methods (Dholakia *et al.*, 2004; Anagnostopoulos *et al.*, 2008; Eagle *et al.*, 2009; Tang *et al.*, 2009; Onnela and Reed-Tsochas, 2010) focus on the problem with a known structure of the network. Tsur and Rappoport (2012) predicted the actions of social media using various content and topology features. Bhagat *et al.* (2012) identified the users with a strong influence based on the experiences and opinions of their friends. Bonchi *et al.* (2011) mined the business applications by learning to combine the community structure and the network dynamics. Saito *et al.* (2011) proposed a probabilistic model based on the observed diffusion data to make the prediction. Wu *et al.* (2011) discovered the behavior patterns of the users by media communications research in Twitter. Bakshy *et al.* (2011) studied the users' actions based on a family of hypothetical marketing strategies. Cha *et al.* (2010) developed three different evaluation measures to learn the users' shopping influence. Some other methods focus on the problem with a network of unobserved links. Gomez-Rodriguez *et al.* (2010) developed a generative model of cascades to trace the paths of users. They later extended this work using convex optimization (Gomez-Rodriguez and Schölkopf, 2012). Duong *et al.* (2011) solved the problem by building a history-dependent graphical multiagent model. Yang and Leskovec (2010) considered users to have the influence functions, which decide the variations of the events in a future time. Zhang *et al.* (2003) applied the correlation analysis to develop a method to discover the links of a network in the application of money laundering crime discovery. Cui *et al.* (2013) predicted the cascading outbreaks of information diffusion by optimizing the variation of logistic regression.

5.2 Mining the users' behavior of social networks

In the field of social network analysis, the research about predicting the users' behavior based on various assumptions of users' situations has generated a large body of the literature. Guille *et al.* (2013) gave the detailed introduction about information diffusion in a social network. Romero *et al.* (2011) considered every user or website for its 'influence' and 'passivity' scores in a social network and then combined the known network structure to discover the activities of users. Myers and Leskovec (2012) considered the relationships of 'cooperation' and 'competition' among the contagions and developed a statistical model to predict the actions of contagions. In their another work (Myers *et al.*, 2012), they also considered that the actions of users are influenced by both internal and external exposures. Anagnostopoulos *et al.* (2011) assumed that in the users' information diffusion, there exist authority users who have higher powers to influence others. Bakshy *et al.* (2012) mined the users' behaviors based on learning two interactions among the users: strong ties and weak ties. Bernstein *et al.* (2013) developed an approach that has the ability to discover the behavior of invisible users in the large-scale log data.

In the context of business model mining and prediction, the existing literature fails to consider the complete relationships among the consumers and among the shops simultaneously. This work stands out by solving this problem through an optimization model using information diffusion theory.

6 Conclusions

We have studied the problem of business model mining and prediction in e-commerce applications. The existing methods fail to consider the potential relationships both among the consumers (consumer influence) and among the shops (competitions or collaborations). Instead of formulating this as a regression or time-series problem, we formulate the problem by combining regression with consumer influence analysis. We then propose a method for e-commerce business model mining and prediction, called EBMM. The challenge that the links in a network are typically not directly observed is addressed by applying information diffusion theory through the consumer–shop network in EBMM. Extensive evaluations using Alibaba Group e-commerce data have demonstrated the superiority of EBMM over the state-of-the-art methods.

References

Anagnostopoulos, A., Kumar, R., Mahdian, M., 2008. Influence and correlation in social networks. Proc. 14th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, p.7-15. [doi:10.1145/1401890.1401897]

Anagnostopoulos, A., Brova, G., Terzi, E., 2011. Peer and authority pressure in information-propagation models. *LNCS*, **6911**:76-91. [doi:10.1007/978-3-642-23780-5_15]

Bakshy, E., Hofman, J.M., Mason, W.A., *et al.*, 2011. Everyone's an influencer: quantifying influence on Twitter. Proc. 4th ACM Int. Conf. on Web Search and Data Mining, p.65-74. [doi:10.1145/1935826.1935845]

Bakshy, E., Rosenn, I., Marlow, C., *et al.*, 2012. The role of social networks in information diffusion. Proc. 21st Int. Conf. on World Wide Web, p.519-528. [doi:10.1145/2187836.2187907]

Bernstein, M.S., Bakshy, E., Burke, M., *et al.*, 2013. Quantifying the invisible audience in social networks. Proc. SIGCHI Conf. on Human Factors in Computing Systems, p.21-30. [doi:10.1145/2470654.2470658]

Bhagat, S., Goyal, A., Lakshmanan, L.V.S., 2012. Maximizing product adoption in social networks. Proc. 5th ACM Int. Conf. on Web Search and Data Mining, p.603-612. [doi:10.1145/2124295.2124368]

Bonchi, F., Castillo, C., Gionis, A., *et al.*, 2011. Social network analysis and mining for business applications. *ACM Trans. Intell. Syst. Technol.*, **2**(3), Article 22. [doi:10.1145/1961189.1961194]

Box, G.E.P., 2008. Time Series Analysis: Forecasting and Control. Wiley. [doi:10.1002/9781118619193]

Boyd, S., Parikh, N., Chu, E., *et al.*, 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, **3**(1):1-122. [doi:10.1561/2200000016]

Cha, M., Haddadi, H., Benevenuto, F., *et al.*, 2010. Measuring user influence in Twitter: the million follower fallacy. Proc. 4th Int. AAAI Conf. on Weblogs and Social Media, p.10-17.

Cui, P., Jin, S.F., Yu, L.Y., *et al.*, 2013. Cascading outbreak prediction in networks: a data-driven approach. Proc. 19th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, p.901-909. [doi:10.1145/2487575.2487639]

Dholakia, U.M., Bagozzi, R.P., Pearo, L.K., 2004. A social influence model of consumer participation in network- and small-group-based virtual communities. *Int. J. Res. Market.*, **21**(3):241-263. [doi:10.1016/j.ijresmar.2003.12.004]

Donoho, D.L., Johnstone, I.M., 1995. Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Statist. Assoc.*, **90**(432):1200-1224. [doi:10.1080/01621459.1995.10476626]

Duong, Q., Wellman, M.P., Singh, S.P., 2011. Modeling information diffusion in networks with unobserved links. SocialCom/PASSAT, p.362-369.

Eagle, N., Pentland, A., Lazer, D., 2009. Inferring friendship network structure by using mobile phone data. *PNAS*, **106**(36):15274-15278. [doi:10.1073/pnas.0900282106]

Friedman, J., Hastie, T., Höfling, H., *et al.*, 2007. Pathwise coordinate optimization. *Ann. Appl. Statist.*, **1**(2):302-332. [doi:10.1214/07-AOAS131]

Gomez-Rodriguez, M., Schölkopf, B., 2012. Influence maximization in continuous time diffusion networks. Int. Conf. on Machine Learning.

Gomez-Rodriguez, M., Leskovec, J., Krause, A., 2010. Inferring networks of diffusion and influence. Proc. 16th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, p.1019-1028. [doi:10.1145/1835804.1835933]

Guille, A., Hacid, H., Favre, C., *et al.*, 2013. Information diffusion in online social networks: a survey. *ACM SIGMOD Rec.*, **42**(1):17-28. [doi:10.1145/2503792.2503797]

Hoefling, H., 2010. A path algorithm for the fused lasso signal approximator. *J. Comput. Graph. Statist.*, **19**(4):984-1006. [doi:10.1198/jcgs.2010.09208]

Long, B., Zhang, Z.F., Yu, P.S., 2007. A probabilistic framework for relational clustering. Proc. 13th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, p.470-479. [doi:10.1145/1281192.1281244]

Myers, S.A., Leskovec, J., 2012. Clash of the contagions: cooperation and competition in information diffusion. IEEE 12th Int. Conf. on Data Mining, p.539-548. [doi:10.1109/ICDM.2012.159]

Myers, S.A., Zhu, C.G., Leskovec, J., 2012. Information diffusion and external influence in networks. Proc. 18th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, p.33-41. [doi:10.1145/2339530.2339540]

Onnela, J.P., Reed-Tsochas, F., 2010. Spontaneous emergence of social influence in online systems. *PNAS*, **107**(43):18375-18380. [doi:10.1073/pnas.0914572107]

Romero, D.M., Galuba, W., Asur, S., *et al.*, 2011. Influence and passivity in social media. European Conf. on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, p.18-33. [doi:10.1007/978-3-642-23808-6_2]

Saito, K., Ohara, K., Yamagishi, Y., *et al.*, 2011. Learning diffusion probability based on node attributes in social networks. 19th Int. Symp. on Foundations of Intelligent Systems, p.153-162. [doi:10.1007/978-3-642-21916-0_18]

Tang, J., Sun, J.M., Wang, C., *et al.*, 2009. Social influence analysis in large-scale networks. Proc. 15th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, p.807-816. [doi:10.1145/1557019.1557108]

Tibshirani, R., Saunders, M., Rosset, S., *et al.*, 2005. Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B*, **67**(1):91-108. [doi:10.1111/j.1467-9868.2005.00490.x]

Tsur, O., Rappoport, A., 2012. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. Proc. 5th ACM Int. Conf. on Web Search and Data Mining, p.643-652. [doi:10.1145/2124295.2124320]

Wu, S.M., Hofman, J.M., Mason, W.A., *et al.*, 2011. Who says what to whom on Twitter. Proc. 20th Int. Conf. on World Wide Web, p.705-714. [doi:10.1145/1963405.1963504]

Yang, J., Leskovec, J., 2010. Modeling information diffusion in implicit networks. IEEE 10th Int. Conf. on Data Mining, p.599-608.

Zhang, Z.F., Salerno, J.J., Yu, P.S., 2003. Applying data mining in investigating money laundering crimes. Proc. 9th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, p.747-752. [doi:10.1145/956750.956851]

Zhongfei ZHANG, editorial board member of *Frontiers of Information Technology & Electronic Engineering*, is a Qiushi Professor at Department of Information Science and Electronic Engineering, Zhejiang University, China. He was on leave from State University of New York (SUNY) at Binghamton, USA, where he was a professor of computer science, and director of the Multimedia Research Laboratory. He received a B.S. in electronics engineering (with honors), an M.S. in information sciences, both from Zhejiang University, China, and a Ph.D. in computer science from the University of Massachusetts at Amherst, USA. He was on the faculty of Computer Science and Engineering Department, and a research scientist at the Center of Excellence for Document Analysis and Recognition, both at SUNY Buffalo, before he joined the faculty of computer science at SUNY Binghamton. His research focuses on data mining and knowledge discovery, specifically on multimedia data mining (e.g., concept discovery from large-scale imagery/video data), relational data mining (e.g., community discovery from large-scale relational data), and general pattern recognition.