Orig-research

Antispam

Collective Hyping Detection System for Identifying Online Spam Activities

Qinzhe Zhang
University of Technology Sydney

Jia Wu
Macquarie University

Peng Zhang, Guodong Long, and Chengqi Zhang
University of Technology Sydney

A new solution aims to identify spam comments and detect products that adopt an evolving spam strategy for promotion. Specifically, an unsupervised learning model combines heterogeneous product review networks to discover collective hyping activities.

Positive online comments about products and stores are very valuable to individuals and online businesses, enticing some merchants to seek and pay for fake reviews from online services to unfairly hype themselves or denigrate their competitors.

Many approaches have been successfully developed to detect online spam. Fangtao Li[1] initially analyzed several attributes related to spam behavior, such as content, sentiment, product, and metadata features, and exploited a two-view semisupervised method to identify spam reviews. Song Feng[2] defined three types of reviewers (any-time, multi-time, and single-time reviewers) and statistically made distributional footprints of deceptive reviews by using neuro-linguistic programming (NLP) techniques. Geli Fei[3] proposed a model to detect spammed products or product groups by comparing the differences in rating behaviors between suspicious and normal users. All these models rely on content features that can be easily found by inserting special characters, but other features, such as temporal and network information, have been employed as well. Qian Xu[4] collected large-scale real-world datasets from telecommunication service providers and combined temporal and user network information to classify spammers using Short Message Service (SMS). Sihong Xie[5] proposed a model that only uses temporal features, with no semantic or rating behavior analysis, to detect abnormal bursts as the number of reviews increases. Finally, Tyler Moore[6] studied the problem of temporal correlations between spam and phishing websites. Intuitively, these works can also be used to uncover sophisticated spam strategies.

Figure 1. Evolving marketing hyping ecosystem.

Amazon has sued more than 1,000 product review sellers who sell fake promotions on Fiverr.com (one of the most famous being Spam Reviewer Cloud; http://money.cnn.com/2015/10/18/technology/amazon-lawsuit-fake-reviews). On such user cloud platforms, business owners can purchase anonymous comments generated by real users by paying for them. It makes spam detection very challenging, as the advent of a massive number of apparently genuine fake reviewers (which we refer to as "genuine fakes" in this article) makes the fraud pattern much more nebulous to track. To date, as Figure 1 shows, many third-party platforms have created various fake review markets for online product sellers and fake review providers. In real-world business processes, massive numbers of random but genuine fake review providers conduct real transactions and write positive comments to claim a bonus (many e-commerce websites think they can reduce spam reviews by allowing only real buyers to write them). Existing research ignores the latent connections in product networks, which are difficult to discover, especially when these spam activities have become a hyping and advertising investment that has gained increased popularity among homogeneous competitors online. Thus, antispam rules can be easily avoided, which also impairs the efficiency and effectiveness of detection performance.

In this work, we coin a new solution—collaborative marketing hyping detection—that aims to detect groups of online stores that simultaneously adopt marketing hyping. This field involves various challenges:

- How can heterogeneous product information network be defined to infer their latent collaborative hyping behaviors? Network information might not be directly observed in the original datasets, so we need to build up a relationship matrix between products to represent their underlying correlation.
- What features need to be selected to best solve our problem? Traditional features such as semantic clues or user relations might no longer be suitable for discovering fraud due to rapidly evolving spam strategies. Hence, we need to choose dedicated features according to our specific scenario.
- How can we design a model that effectively identifies collaborative marketing hyping behavior? A model that can employ the power of heterogeneous product networks to discover collective hyping behavior is required here.

To overcome these challenges, we propose an unsupervised shapelet learning model to discover the temporal features of product reviews and then integrate the heterogeneous product network information as regularization terms, to discover the products that are subject to collaborative hyping. We define three regularization terms that reflect the underlying correlations among users, products, and online store networks.

Problem Definition

In 2015, fake product issues on Taobao were exposed on many public and social media platforms. Official investigation, conducted by the Chinese Consumer Association (CCA), found that most of the fake products surprisingly maintained a top ranking position, which could continuously damage consumer interests. A key factor here is that only individuals who successfully purchase a product can leave comments on that product on Taobao. CCA also reported that several notorious fake review web markets in China had formed a fake review chain. On such a platform, for example, a Taobao store owner can post a request, say, for 1,000 reviews at 10 RMB each, as 1,000 tasks. Anyone in China who has the time can earn 10 RMB if they know such a web market and can write a review.

Ironically, these platform providers have their own mechanisms for preventing people from spamming the tasks, not only guaranteeing that a person can take only one task posted by a specific store, but they can also ensure that the least amount of spam evidence (semantic clues, user behavior, and so on) is left in the comments, thereby subtly escaping most traditional spam detection rules. These stores usually purchase fake reviews periodically, as individual needs change over time. For instance, by predicting the most active shopping periods for individuals— festivals, end or beginning season sales, and so on— online merchants buy fake reviews months in advance to hold the top position until people start purchasing for real, to drive business to their sites. This forms a collaborative marketing hyping phenomenon among all homogeneous brands and disadvantages honest shop owners.

Mathematically, let's consider a set of online products $\mathbf{P}$ belonging to a group of stores; each product $\mathbf{p}$ in $\mathbf{P}$, a review time series $\mathbf{t}$ can be obtained. Consider a set of time series $\mathbf{T} = \{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_n\}$ that correspond to $\mathbf{P}$. Each time series $\mathbf{t}_i$ $(1 \le i \le n)$ contains an ordered set of real values denoted as $(\mathbf{t}_{i(1)}, \mathbf{t}_{i(2)}, \ldots, \mathbf{t}_{i(q_i)})$, where q_i is the length of $\mathbf{t}_i$. We wish to learn a set of top-k most discriminative shapelets $\mathbf{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_k\}$. Similar to the shapelet learning model,[7] we set the length of the shapelets to expand r different length scales starting at a minimum l_{min}, that is, $\{l_{min}, 2 \times l_{min}, \ldots, r \times l_{min}\}\}$. Each length scale $i \times l_{min}$ contains k_i shapelets and $k = \sum_{i=1}^{r} k_i$. Clearly, $\mathbf{S} \in \bigcup_{i=1}^{r} \mathbf{R}^{k_i \times (i \times l_{min})}$ and $r \times l_{min} \ll q_i$ to keep the shapelets compact. Our shapelet learning model uses matrix factorization techniques that we discuss later; all products can be classified on latent spaces according to their temporal features. Additionally, we don't only consider singular spam activities in one store but aim to detect collaborative hyping behaviors, thus, three different product information networks are defined as regularization terms to constrain matrix factorization. Table 1 summarizes the symbols and notations used in this article.

Methodology

Before we go much further, let's first discuss the shapelet learning model and the product information network for regularization. Then we can formulate the objective function.

Shapelet Learning Model

Shapelets are discriminative subsequences of time series that best predict the target variable, while shapelet learning models are usually designed with a classification purpose that aims to identify the similarity between two items refer to their temporal feature. **Shapelet-transformed representation.** According to Jason Lines's[8] work, *shapelet transformation* was proposed to downsize the time series into a short feature vector in the shapelet feature space. Time series are orderless but can be uniformly represented by shapelet transformation that preserve the most relevant information for classification.

Table 1. Symbols and notations used in this article.

O	Series of target online stores
P	Group of products belonging to **O**
T	Time-series dataset generated from **P**
S	Top-k most discriminative shapelets
X	Shapelet transformation matrix
d_{ij}	Distance between shapelets i and j
l_i	Shapelet length
q_j	Time-series length

E	Shapelet similarity matrix
V	Pseudo-class label matrix
U	Classification boundary under **V**
G_1	Store-based network
G_2	Product-based network
G_3	User correlation-based network
R(*SBR*)	Store-based regularization
R(*PBR*)	Product-based regularization
R(*UCR*)	User correlation-based regularization

For instance, given a set of time series $\mathbf{T} = \{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_n\}$ and a set of shapelets $\mathbf{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_k\}$, we use $\mathbf{X} \in \mathbf{R}^{k \times n}$ to denote the shapelet-transformed matrix, where each element $\mathbf{X}_{\mathbf{s}_i \mathbf{t}_j}$ denotes the distance between shapelet $\mathbf{s}_i$ and time series $\mathbf{t}_j$. We use $\mathbf{X}_{(ij)}$ to represent $\mathbf{X}_{(\mathbf{s}_i, \mathbf{t}_j)}$ which can be calculated as

$$\mathbf{X}_{(ij)} = \min_{g=1,\ldots,\bar{q}} \frac{1}{l_i} \sum_{h=1} (\mathbf{t}_{j(g+h\_1)} - \mathbf{s}_{i(h)}), \qquad (1)$$

//Please note: I couldn't make out what was at the top of this equation.//

where $q = q_j - l_i + 1$ is quantity of segments with length l_i from series $\mathbf{t}_j$, and q_j, l_i are the lengths of time series $\mathbf{t}_j$ and shapelet $\mathbf{s}_i$, respectively.

Given a set of time-series data $\mathbf{S}$, $\mathbf{X}_{(ij)}$ is a function that refers to all candidate shapelets $\mathbf{S}$, that is, $\mathbf{X}(\mathbf{S})_{(ij)}$. Here, we elide the variable $\mathbf{S}$ and use $\mathbf{X}_{(ij)}$ instead.

Based on Lines,[7] we approximate the distance function using the *soft minimum function* as in Equation 2:

$$\mathbf{X}_{(ij)} \approx \frac{\sum_{q=1}^{L_q} d_{ijq} \cdot e^{\alpha d_{ijq}}}{\sum_{q=1}^{L_{\bar{q}}} e^{\alpha d_{ijq}}}, \qquad (2)$$

where $d_{ijq} = \frac{1}{\bar{l}_i} \sum_{h=1}^{L_{l_i}} (\mathbf{t}_{j(q+h1)} - \mathbf{s}_{i(h)})$ and α control function precision. The soft minimum approaches the true minimum when $\alpha \to -\infty$. In our experiments, we set $\alpha = -100$.

Pseudo-class label. Our dataset has been human labeled, so to evaluate our model's accuracy, we introduce *pseudo-class labels* for unsupervised learning. In this article, c denotes the number of pseudo classes. The pseudo-class label matrix $\mathbf{V} \in \mathbf{R}^{c \times n}$ contains c labels, where $\mathbf{V}_{(ij)}$ indicates the probability of the jth time-series candidate categorized into the ith class. If $V_{(\bar{i}j)} > V_{(i,j)}, \forall i$, then the time-series example $\mathbf{t}_j$ belongs to the cluster i.

Shapelet similarity minimization. To maximize the variance of shapelets, we penalize the model if similar shapelets are generated. We denote the shapelet similarity matrix as $\mathbf{E} \in \mathbf{R}^{k \times k}$, where each element $\mathbf{E}_{(\mathbf{s}_i \mathbf{s}_j)}$ represents the similarity between two shapelets $\mathbf{s}_i$ and $\mathbf{s}_j$. $\mathbf{E}_{(ij)}$ represents $\mathbf{E}_{(\mathbf{s}_i, \mathbf{s}_j)}$:

$$\mathbf{E}_{(ij)} = e^{\frac{\|d_{ij}\|^2}{\sigma 2}}, \quad (3)$$

where d_{ij} is the distance between shapelet $\mathbf{s}_i$ and shapelet $\mathbf{s}_j$; it can be calculated by following Equation 2.

Shapelet learning model. We measure the least-square error between the original shapelet transformation matrix X and the pseudo-class labels by minimizing their distance as below:

$$\min \|\mathbf{X} - \mathbf{UV}\|_{F}^{2}, \quad (4)$$

where $\mathbf{U} \in R^{k \times c}$ is classification boundary corresponding to pseudo-class labels V. Overall, this is a joint optimization problem with respect to variables S, U, and V for this model, as in Equation 5:

$$\min_{\mathbf{S},\mathbf{U},\mathbf{V}} \frac{1}{2}\|\mathbf{X} - \mathbf{UV}\|_F^2 + \frac{\lambda_3}{2}\|\mathbf{H(s)}\|_F^2 + \frac{\lambda_4}{2}\|\mathbf{U}\|_F^2, \quad (5)$$

Product Network Regularization

The product network provides correlation information about all online stores. We model three types of heterogeneous information network as regularization terms: store-based regularization, product-based regularization, and user-correlation regularization.

Store-based regularization. Online sellers typically hype their products periodically to retain their top ranking position. Homogeneous competitors always observe the hyping action of their peers when they prepare to purchase fake reviews. Collaborative marketing hyping is essentially a new type of ranking position competition between homogeneous stores and product owners: for instance, store A, which sells protein powder, will start to hype when it finds that its competitor has begun to seek spammers. Thus, similar products belonging to different stores could share common hyping behaviors in terms of their temporal features. Based on the above analysis, we can design a store-based regularization term $R(SBR)$ to connect all the same products within different stores:

$$R(SBR) = \mathbf{VG}_1\mathbf{V}^{\mathrm{T}}. \quad (6)$$

where $\mathbf{G}_1$ is a store-based network matrix.

As can be seen in Figure 2, we set up the connection values as 1 for every product belonging to the same merchants; otherwise, it's 0. Store-based regularization terms based on the assumption of homogeneous merchandise within different stores share a similar hyping pattern with respect to their time-series features. Thus, there's a very large possibility that they'll be categorized into the same cluster. However, not all high-ranking products enhance reputation and profit by adopting spam method.

Figure 2. Store-based network and matrix $\mathbf{G}_1$.

Product-based regularization. In contrast to the external comparison in store-based regularization, product-based regularization focuses on the internal comparison of different products within the same store. An online seller who decides to use fake reviews won't hype only a single product in their stores, hence, we introduce the product-based regularization term $R(PBR)$ to indicate homogeneous competitor products within different stores:

$$R(PBR) = \mathbf{V}\mathbf{G}_2\mathbf{V}^{\mathrm{T}}, \qquad (7)$$

where $\mathbf{G}_0$ is the product connection matrix and $G_{ij} = 1$ when these two products i and j are within the same store; otherwise, $G_{ij} = 0$. Intuitively, those merchants will adopt unfair techniques to promote most of their products, rather than only hyping one or two of them. Such products can also be more likely to share similar temporal patterns, so this step is thus an ideal supplement to the first regularization model. Figure 3 describes the product-based network and matrix.

Figure 3. Product-based network and matrix $\mathbf{G}_2$.

Figure 4. User correlation-based network and matrix $\mathbf{G}_3$.

User correlation-based regularization. Spammers can accept multiple fake review tasks corresponding to different products at the same time. Thus, during the positive review burst period, those spammers can simultaneously emerge in the review list of hyping-oriented products. Ordinary customers don't normally purchase the same product from different stores at the same time, nor do they buy different products in different stores (to save transportation costs). Hence, we introduce the user correlation-based regularization term $R(UCR)$ to minimize the difference between products reviewed by same user in specified period:

$$R(UCR) = \mathbf{V}\mathbf{G}_3\mathbf{V}^{\mathrm{T}}, \qquad (8)$$

where $G_{ij} = 1$ in $\mathbf{G}_3$ when these two products i and j are reviewed by the same users in a nominated period; otherwise, $G_{ij} = 0$. Product network information based on the evidence of spammer groups is also very important for avoiding information loss; Figure 4 shows the user-correlation product network.

Collaborative Hyping Detection Model

We propose our collaborative hyping detection model (CHDM) to solve the collective marketing hyping problem defined earlier. This model integrates all the regularization terms we've defined into a shapelet learning model that utilizes temporal features and product network information for clustering. The objective function is given as

$$\min_{\mathbf{S},\mathbf{U},\mathbf{V}} \frac{1}{2}\|\mathbf{X}-\mathbf{U}\mathbf{V}\|_F^2 + \frac{\lambda_1}{2}\|\mathbf{V}\mathbf{G}_1\mathbf{V}^{\mathrm{T}}\| + \frac{\lambda_2}{2}\|\mathbf{V}\mathbf{G}_2\mathbf{V}^{\mathrm{T}}\| + \frac{\lambda_3}{2}\|\mathbf{V}\mathbf{G}_3\mathbf{V}^{\mathrm{T}}\| + \frac{\lambda_4}{2}\|\mathbf{H}(\mathbf{s})\|_F^2 + \frac{\lambda_5}{2}\|\mathbf{U}\|_F^2. \qquad (9)$$

Figure 5. Framework of the collaborative hyping detection model (CHDM).

Algorithm

Our proposed CHDM is very straightforward and integrates all the network information generated from store, product, and user correlation into a shapelet learning model. Figure 5 shows our model's framework. Specifically, we employ an unsupervised learning model to cluster the target stores based on their comments' temporal features. In addition, we incorporate the three different network information (store, product, and user correlation) as regularization terms to enhance clustering accuracy.

To achieve a local minimum of the CHDM objective function given by Equation 9, we conduct the coordinate gradient descent to iteratively solve the three variables as in Algorithm 1.

//start algorithm//

Input: *T*, review sequential data;

c, number of class;

l_{min}, *k*, length & number of sequential features;

i_{max}, number of internal iterations;

η, the learning rate;

λ_1, λ_2, λ_3, λ_4, λ_5, and α, σ, parameters

Output: Sequential feature **S** and class label **V**

Initialize $\mathbf{S}_0$, $\mathbf{V}_0$, $\mathbf{U}_0$

while Not convergent **do**

1. *UpdateVwithFixedUandS* :

$$\mathbf{V}_{ij}^{i+1} = \mathbf{V}_{ij}^{t} \frac{(X_t^T U_t)_{ij}}{\left[(\lambda_1 G^T + \lambda_2 G_2^T + \lambda_3 G_3^T + V_t V_t^T) X_t^T U_t\right]_{ij}}$$

2. *UpdateVwithFixedVandS* :

$\mathbf{U}^{i+1}=$ **//I couldn't make out this equation//**

3. *UpdateVwithFixedUandV* ::

$$\mathbf{S}^{t+1} = \mathbf{S}^{t} - \alpha[(\mathbf{X}_s - \mathbf{UV})\frac{\alpha \mathbf{X}_s}{\alpha \mathbf{S}} + \mathbf{H}_s \frac{\alpha \mathbf{H}_s}{\alpha \mathbf{S}}]$$

Output $S^* = S_{t+1}$; $U^* = U_{t+1}$; $V^* = V_{t+1}$

Algorithm 1. CHDM algorithm.

//end algorithm//

Experiments

We validated our method by attempting to answer two questions: How well does CHDM outperform other state-of-the-art spam detection techniques that also utilize temporal features? What's the respective contribution of each of the defined regularization terms (PBR, SBR, and UCR) to our proposed model?

Dataset

The counterfeit crisis on Taobao caused a stir in 2015 when the CCA reported the top 10 fake goods sellers on Taobao, which included clothing, makeup, and digital devices, among others.

Accordingly, we collected product data from stores in these industries. It should be noted that our goal isn't to detect fake products, but those that reported high-ranking products are all very susceptible to hyping. Table 2 describes the statistics of our real-world dataset.

User Correlation-Based Product Network

User name is a key piece of evidence for recognizing users, but user information in Taobao is anonymized and IDs can't be acquired, so we can only use an approximate match method to identify spammers. For instance, a user name on a review page might appear as "D***d," which indicates that only the initial and last characters in the name were kept. By matching the characters in these two position, we can at least approximately identify the same (or similar) users.

Table 2. Dataset statistics

Dataset	Store	Products	Reviews
Clothing products	25	186	215,892
Cosmetics products	22	177	225,823
Electronic products	18	159	209,654
Food products	19	165	208,639
Health products	24	201	248,536
Footwear products	20	199	190,953

The above name evidence is insufficient, however, and could cause noise and inaccuracy, hence we introduce another important piece of evidence— user level. Taobao applies very strict mechanisms in user-level upgrades. Only users who have successfully completed transactions with online shop owners can accumulate the required scores to upgrade to a higher level. The higher the user level, the much more scores are needed to upgrade, thus, as additional information in the user-matching process, it decreases the inaccuracy caused by only using the user name information. We can build up a user correlation-based product network by following these steps:

1. For a pair of randomly selected products, we first observe their review burst period and put all related reviewers into two separate lists.
2. We then conduct a matching process, using reviewers' names as well as their user level and place-matched users on a short list
3. If the shortlist contains more than five matched users, we set the connection value of these two products to 1 in the user correlation-based product network matrix. Otherwise, the value is set to 0.
4. Iteratively, we match all the products in our dataset to build a user correlation-based product network.

Experiment Results

In this section, we first discuss the comparison results between our model and other two benchmark methods, then we discuss the parameter study by following two case studies.

Comparison with existing models. In our experiment, we invited 20 experienced online buyers with high user levels to label products according to their evaluation. Some of these invited users are domain experts who have previously written fake reviews for shops.

To validate our proposed CHDM, we compare it with two representative spam detection techniques that mainly utilize temporal features:

- Multiscale spam detection (MSSD)[5] detects singleton reviewers who appear in an assigned time window as abnormal evidence of spam activities.
- Collective hyping spam detection (CHSD)[9] employs a temporal feature classification technique with product-related network information for collective spam detection.

Table 3 clearly shows that our method significantly outperforms the two baseline techniques, and we can make the following related observations:

- Our CHDM can achieve about 5 to 10 percent higher accuracy than CHSD and MSSD models, respectively.
- The spam temporal feature is implicit, and the significant drop in MSSD indicates that latent information can't be fully discovered in the human-assigned time window.

Table 3. Comparisons with benchmark method.

	Precision	*Recall*	*Accuracy*
MSSD	0.872 ± 0.0086	0.834 ± 0.0107	0.826 ± 0.0093
CHSD	0.938 ± 0.0082	0.903 ± 0.0059	0.898 ± 0.0064
CHDM	**0.966 ± 0.0091**	**0.950 ± 0.0077**	**0.945 ± 0.0085**

The impact of PBR/SBR/UCR in CHDM. By setting parameters λ_1, λ_2, or λ_3 in our objective function, regularization term integration can be categorized as below:

- $CHDM_{NoReg}$ doesn't consider any regularization terms by setting λ_1, λ_2, and λ_3 to 0.
- $CHDM_S$ considers only store-based regularization terms by setting λ_2 and λ_3 to 0.
- $CHDM_P$ considers only product-based regularization terms by setting λ_1 and λ_3 to 0.
- $CHDM_U$ considers only user correlation-based regularization terms by setting λ_1 and λ_2 to 0.
- $CHDM_{P+S}$ considers store- and product-based regularization terms by setting λ_3 to 0, which is also equal to the CHSD model.[9]
- $CHDM_{S+U}$ considers store- and user correlation-based regularization terms by setting λ_2 to 0.
- $CHDM_{P+U}$ considers product- and user correlation-based regularization terms by setting λ_1 to 0.
- $CHDM$ consider all the regularization terms by setting a set of best-fitting parameters.

Table 4 shows that $CHDM_{NoReg}$ returns the worst result, whereas CHDM outperforms all its counterparts. Additionally, by comparing two other singular regularization integrated models ($CHDM_S$ and $CHDM_P$), we can observe a slight enhancement in $CHDM_U$. This indicates that user correlation-based information is very important, which can also be observed in the comparison of the dual regularization integration models, $CHDM_{P+S}$, $CHDM_{P+U}$, and $CHDM_{S+U}$. In general, these observations prove the existence of collaborative hyping activities in Taobao, and our model successfully takes the spam reviewer correlation into account to find the products involved in collective spamming.

Case study. As previously discussed, the MSSD model[5] identifies spam stores or products one by one by detecting abnormal singleton reviewers appearing in an assigned time window. However, this method misses the latent information that underlies evolving hyping activities. We pick up two of the representative cases in Figure 6, which were tagged as "spam" by the MSSD model but that our model placed in a "clean" class. Apparently, there's a remarkable purchasing burst in both of them, with 80 percent of buyers in this time window being singleton reviewers. In our experiment, we define customers who have made less than five transactions online since their registration as singleton reviewers. Because of the different customer level-segmentation strategies and privacy policies in Taobao, this provides the best match with the definition of singleton reviewers in the MSSD model.

Figure 6. Case study analysis.

Table 4. Comparisons with benchmark method.

	Precision	Recall	Accuracy
$CHDM_{NoReg}$	0.903 ± 0.0121	0.874 ± 0.0107	0.867 ± 0.0082
$CHDM_S$	0.915 ± 0.0029	0.887 ± 0.0059	0.880 ± 0.0117
$CHDM_P$	0.917 ± 0.0038	0.890 ± 0.0114	0.883 ± 0.0036
$CHDM_U$	0.924 ± 0.0119	0.901 ± 0.0065	0.894 ± 0.0098
$CHDM_{P+S}$	0.938 ± 0.0082	0.903 ± 0.0059	0.898 ± 0.0064
$CHDM_{S+U}$	0.945 ± 0.0067	0.919 ± 0.0018	0.915 ± 0.0049
$CHDM_{P+U}$	0.953 ± 0.0074	0.928 ± 0.0063	0.922 ± 0.0108
CHDM	**0.966 ± 0.0091**	**0.950 ± 0.0077**	**0.945 ± 0.0085**

Figure 7. Collaborative marketing hyping activities.

To validate, we asked domain experts to recheck these two cases; most are of the opinion that it's a normal situation as one of the burst periods is close to Christmas while another is close to "Double 11" (online Boxing Day in China hyped by Taobao).

Collaborative marketing hyping activities. For each spam-involved industry exposed by the CCA, we picked several examples to demonstrate collaborative marketing hyping activities (see Figure 7). For instance, in healthy product industries, CHDM recognizes similar temporal patterns between May and June 2015. We can clearly observe a gradually ascending curve in terms of sales volume for three different products, which means that store owners no longer adopt the previous kinds of abrupt spam strategies to escape the detection algorithm applied by Taobao. They gradually increase the number of hyping purchasers at the beginning of May so that their product is ranked in the top position by Mother's Day in the middle of May or June 2015 or Father's Day in June 2015 by carefully fitting the Taobao ranking algorithm. A similar situation can be observed in other industries, verifying that our model can successfully identify collaborative marketing hyping activities.

There are several directions to be explored in the future. Sematic information hasn't been taken into consideration, so it will be interesting to combine this type of data into our model. Moreover, we can employ more pieces of information to match users, such as location or review sentiment analysis, to help detect spam groups.

Acknowledgments

This work was supported by Australian Research Council (ARC) Discovery Projects (nos. DP140102206 and DP140100545) and Linkage Projects (nos. LP150100671 and LP160100630).

References

1. F. Li, M. Huang, Y. Yang, and X. Zhu, "Learning to Identify Review Spam," *Proc. Int'l Joint Conf. Artificial Intelligence*, 2011, pp. 2488–2493.

2. S. Feng et al., "Distributional Footprints of Deceptive Product Reviews," *Proc. Int'l Conf. Web and Social Media*, 2012, pp. 98–105.

3. G. Fei et al., "Exploiting Burstiness in Reviews for Review Spammer Detection," *Proc. Int'l Conf. Web and Social Media*, 2013, pp. 175–184.

4. Q. Xu et al., "SMS Spam Detection Using Noncontent Features," *IEEE Intelligent Systems*, vol. 27, no. 6, 2012, pp. 44–51.

5. S. Xie et al., "Review Spam Detection via Temporal Pattern Discovery," *Proc. ACM Int'l Conf. Knowledge Discovery and Data Mining*, 2012, pp. 823–831.

6. T. Moore, R. Clayton, and H. Stern, "Temporal Correlations between Spam and Phishing Websites," *Proc. 2nd Usenix Conf. Large-scale Exploits and Emergent Threats*, 2009, p. 5.

7. J. Grabocka et al., "Learning Time-Series Shapelets," P*roc. ACM Int'l Conf. Knowledge Discovery and Data Mining*, 2014, pp. 392–401.

8. J. Lines et al., "A Shapelet Transform for Time Series Classification," *Proc. ACM Int'l Conf. Knowledge Discovery and Data Mining*, 2012, pp. 289–297.

9. Q. Zhang et al., "Exploring Heterogeneous Product Networks for Discovering Collective Marketing Hyping Behavior," *Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining*, 2016, pp. 40–51.

Qinzhe Zhang is a PhD candidate in computer science at the University of Technology Sydney. His research interests include data mining, machine learning, social recommendation, online spam detection, and artificial intelligence. Contact him at Qinzhe.Zhang@student.uts.edu.au.

Jia Wu, corresponding author, is a lecturer in the Department of Computing, Faculty of Science and Engineering, at Macquarie University, Sydney, Australia. His research focuses on data mining and machine learning. Wu received a PhD in computer science from the University of Technology Sydney, Australia. He's a member of IEEE. Contact him at Jia.Wu@mq.edu.au.

Peng Zhang is a lecturer in the Centre of Artificial Intelligence at the University of Technology Sydney. He received aa PhD in computer science from the Chinese Academy of Sciences. Zhang won the Best Paper Award at ICCS-14. Contact him at Peng.Zhang@uts.edu.au.

Guodong Long is a lecturer in the Centre for Artificial Intelligence at the University of Technology Sydney.

His research interests include data mining, machine learning, artificial intelligence, and cloud computing. Long received a PhD in information technology from the University of Technology Sydney. Contact him at Guodong.Long@uts.edu.au.

Chengqi Zhang is a professor of information technology at the University of Technology Sydney and director of its Priority Investment Research Centre for Quantum Computation and Intelligent Systems. His research interests mainly focus on data mining and its applications. Zhang received a PhD in computer science from the University of Queensland and a D.Sc. (higher doctorate) in computer science from Deakin University. He's a Fellow of the Australian Computer Society and a Senior Member of IEEE. Contact him at Chengqi.Zhang@uts.edu.au.