Latent Topic Analysis for Predicting Group Purchasing Behavior on the Social Web

Feng-Tso Sun, Martin Griss, and Ole Mengshoel
Electrical and Computer Engineering Department
Carnegie Mellon University

Yi-Ting Yeh
Computer Science Department
Stanford University

Abstract

Group-deal websites, where customers purchase products or services in groups, are an interesting phenomenon on the Web. Each purchase is kicked off by a group initiator, and other customers can join in. Customers form communities with people with similar interests and preferences (as in a social network), and this drives bulk purchasing (similar to online stores, but in larger quantities per order, thus customers get a better deal). In this work, we aim to better understand what factors influence customers' purchasing behavior for such social group-deal websites. We propose two probabilistic graphical models, i.e., a product-centric inference model (PCIM) and a group-initiator-centric inference model (GICIM), based on Latent Dirichlet Allocation (LDA). Instead of merely using customers' own purchase history to predict purchasing decisions, these two models include other social factors. Using a lift curve analysis, we show that by including social factors in the inference models, PCIM achieves 35% of the target customers within 5% of the total number of customers while GICIM is able to reach 85% of the target customers. Both PCIM and GICIM outperform random guessing and models that do not take social factors into account.

1 Introduction

Group purchasing is a business model that offers various deals-of-the-day and an extra discount depending on the size of the purchasing group. After group-deal websites, such as Groupon and LivingSocial, have gained attention, similar websites, such as **ihergo**[1] and **Taobao**,[2] have introduced social networks as a feature for their users. These group-deal websites provide an interesting hybrid of social networks (e.g., Facebook.com and LinkedIn.com) and online stores (e.g., Amazon.com and Buy.com). Customers form communities with people with similar interests and preferences (as in a social network), and this drives bulk purchasing (similar to online stores, but in larger quantities per order, thus customers get a better deal). As we see more and more social interactions among customers in group-deal websites, it is critical to understand the interplay between social factors and purchasing preferences.

In this paper, we analyze a transactional dataset from the largest social group-deal website in Taiwan, ihergo.com. Figure 1 shows a screenshot from the group-deal page of ihergo.com. Each group-purchasing event on ihergo.com consists of three major components: (1) a group initiator, (2) a number of group members, and (3) a group-deal product. A group initiator starts a group-purchasing event for a specific group-deal product. While this event will be posted publicly, the group initiator's friends will also be notified. A user can choose to join the purchasing event to become a group member.

Group initiators play important roles on this kind of group-deal websites. Usually, the merchants would offer incentives for the group initiators to initiate group-purchasing events by giving them products for free if the size of the group exceeds some threshold. In addition, to save shipping costs, the group can choose to have the whole group-deal order shipped to the initiator. In this case, the initiator would need to distribute the products to group members in person. Hence, the group members usually reside or work in the proximity of the group initiator. Sometimes, they are friends or co-workers of the initiator.

Understanding customers' purchasing behavior in this

[1] http://www.ihergo.com

[2] http://www.taobao.com

kind of social group-purchasing scenario could help group-deal websites strategically design their offerings. Traditionally, customers search for or browse products of their interests on websites like Amazon.com. However, on social group-deal websites, customers can perform not only product search, but they can also browse group deals and search for initiators by ratings and locations. Therefore, a good recommender system [1] for social group-deal websites should take this into account. If the website can predict which customers are more likely to join a group-purchasing event started by a specific initiator, it can maximize group sizes and merchants' profits in a shorter period of time by delivering targeted advertising. For example, instead of spamming everyone, the website can send out notifications or coupons to the users who are most likely to join the group-purchasing events.

In this work, we aim to predict potential customers who are most likely to join a group-purchasing event. We apply Latent Dirichlet Allocation (LDA) [2] to capture customers' purchasing preferences, and evaluate our proposed predictive models based on a one-year group-purchasing dataset from ihergo.com.

Our contributions in understanding the importance of social factors for group-deal customers' decisions are the following:

- **A new type of group-purchasing dataset.** We introduce and analyze a new type of group-purchasing dataset, which consists of 5,602 users, 26,619 products and 13,609 group-purchasing events.
- **Predictive models for group-deal customers.** Based on topic models, we propose two predictive models that include social factor. They achieve higher prediction accuracy compared to the baseline models.

In the next section, we describe related work in the area of group purchasing behavior, social recommendations, and topic models for customer preferences. Section 3 introduces and analyzes the characteristics of our real-world group-purchasing dataset. In Section 4, we first review LDA, then present two proposed predictive models for group-deal customer prediction. Experimental results are given in Section 5. Finally, conclusion and future research direction are presented in Section 6.

2 Related Work

In this section, we review related work in three areas: (1) group purchasing behavior, (2) social recommendations, and (3) topic models for customer preferences.

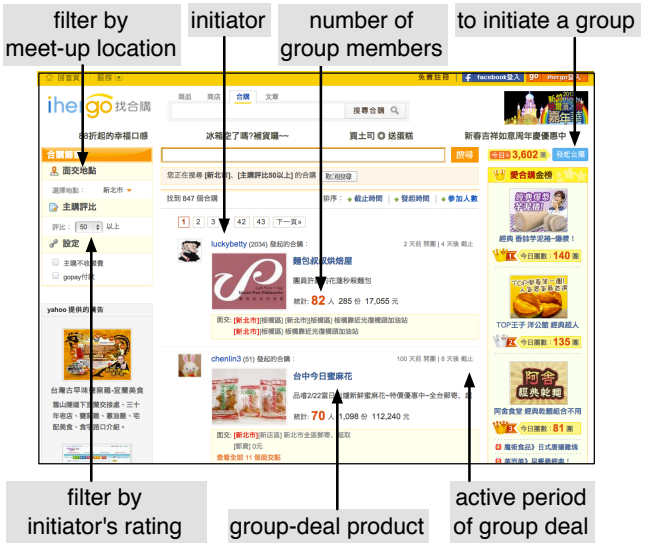


Figure 1: Screenshot of the group-deal page from ihergo.com.

Group Purchasing Behavior. Since group-deal websites such as Groupon and LivingSocial gained attention, several studies have been conducted to understand factors influencing group purchasing behavior. Byers et al. analyzed purchase histories of Groupon and LivingSocial [3]. They showed that Groupon optimizes deal offers strategically by giving "soft" incentives, such as deal scheduling and duration, to encourage purchases. Byers et al. also compared Groupon and LivingSocial sales with additional datasets from Yelp's reviews and Facebook's like counts [4]. They showed that group-deal sites benefit significantly from word-of-mouth effects on users' reviews during sales events. Edelman et al. studied the benefits and drawbacks of using Groupon from the point of view of the merchants [6]. Their work modeled whether advertising and price discrimination effects can make discounts profitable. Ye et al. introduced a predictive dynamic model for group purchasing behavior. This model incorporates social propagation effects to predict the popularity of group deals as a function of time [19]. In this work, we focus on potential customer prediction, as opposed to modeling the overall deal purchasing sales over time.

Social Recommendations. In real life, a customer's purchasing decision is influenced by his or her social ties. Guo et al. analyzed the dataset from the largest Chinese e-commerce website, Taobao, to study the relationship between information passed among buyers and purchasing decision [7]. Leskovec et al. used a stochastic model to explain the propagation of recommendations and cascade sizes [11]. They showed

that social factors have a different level of impact on user purchasing decision for different products. Moreover, previous work also tried to incorporate social information into existing recommendation techniques, such as collaborative filtering [13, 14, 20, 12]. Recently, many recommendation systems have been implemented, taking advantage of social network information in addition to users' preferences to improve recommendation accuracy. For example, Yang et al. proposed a Bayesian-inference based movie recommendation system for online social networks [18]. Our work considers the relationship between the group initiator and the group members as a social tie to augment customer prediction for group-purchasing events.

Topic Models for Customer Preference. Topic models such as LDA have been widely and successfully used in many applications including language modeling [2], text mining [17], human behavior modeling [9], social network analysis [5], and collaborative filtering [8]. Researchers have also proposed new topic models for purchasing behavior modeling. For example, topic models have been extended with price information to analyze purchase data [10]. By estimating the mean and the variance of the price for each product, the proposed model can cluster related items by taking their price ranges into account. Iwata and Watanabe proposed a topic model for tracking time-varying consumer purchase behavior, in which consumer interests and item trends change over time [9]. In this paper, we use LDA to learn topic proportions from purchase history to represent customers' purchasing preferences.

3 Group-Purchasing Dataset

The dataset for our data analysis comes from users' transactional data of a group-deal website, ihergo. It is the largest social group-deal website in Taiwan. We collected longitudinal data between October 1st 2011 and October 1st 2012. From the users' geographical profile, we are able to group them based on their living area. For this study, we include all 5,602 users living in Taipei, the capital of Taiwan. In total, our dataset contains 26,619 products and 13,609 group-purchasing events.

On ihergo, users can purchase a product by joining a group-purchasing event. There are two roles among the users: 1) the **group initiator** and 2) the **group member**. A group initiator initiates a purchase group which other users can join to become group members. Once the group size exceeds some threshold, the group members can get a discount on the product while the initiator can get the product for free. Sometimes the group initiator and the group members already know each other before they join the same group-purchasing event. Sometimes they become friends after the event. Moreover, each user can become a follower of a group initiator. When a new group-purchasing event is initiated by an initiator, the system will notify his or her followers.

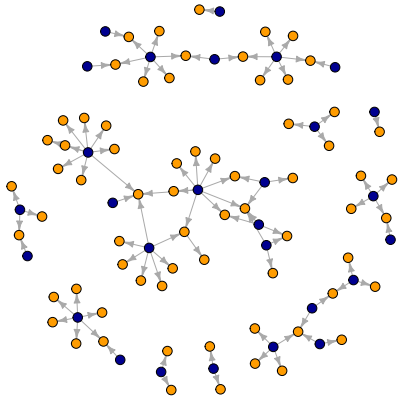


Figure 2: Part of group deal graph for ihergo dataset: illustration of the member-centric relationships between group members and initiators from a subset of randomly sampled joined group members. A directed edge is from a joined customer (dark blue) to an initiator (light orange).

Each group-purchasing deal in our dataset is composed of a set of attributes: the product description (e.g., discounted price, limited quantity, and product category), the group size, the group initiator, the group members, and the time period in which the deal is active. Group-purchasing deals are defined by a time series $\boldsymbol{D} = (D_1, D_2, ..., D_n)$, where D_i is a tuple $(t, p, o, \mathbf{m})$ denoting that a group-purchasing deal for product p is initiated by an organizer (initiator) o with joined group members $\mathbf{m} = \{m_1, ..., m_k\}$ at time t.

We represent group-purchasing events as a directed graph. Each user is a vertex in the graph. For every group-purchasing deal, we build directed edges from each group member to the initiator. There are 5,602 vertices and 16,749 edges in our ihergo dataset. The directed edges are defined by $E = \cup_{i\in[1,n]} \cup_{j\in[1,d(i)]} (m_{i,j}, o_i)$, where $d(i)$ is the number of joined customers for group deal i. The vertices in the graph are defined by $V = M \cup O$, where M denotes all group members $M = \mathbf{m_1} \cup \ldots \cup \mathbf{m_n}$ and O denotes total group-purchasing organizers (initiators) $O = \{o_1\} \cup \ldots \cup \{o_n\}$.

Figure 2 illustrates the joined customer centric graph structure by showing the relationships among a subset of randomly sampled joined customers. Light orange and dark blue vertices represent the group initiators and group members, respectively. According to this dataset, each user has joined 84 group-purchasing

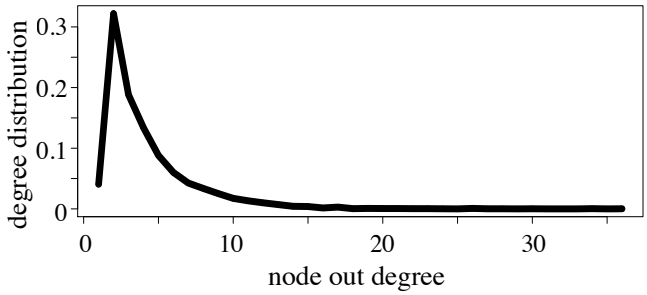


Figure 3: Node out-degree distribution for the group-purchasing graph of ihergo. 80% of the users follow five or fewer initiators.

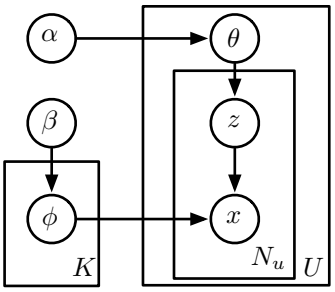


Figure 4: Graphical model representation of the latent Dirichlet allocation model.

events on average. However, one interesting observation from the graph is that the number of outgoing edges from dark blue vertices is far less than 84. This property can be even clearly seen from the out-degree distribution for the overall group-purchasing graph shown in Figure 3. We see that 80% of the users only join group-purchasing events initiated by 5 or fewer different initiators. Group members have a tendency to repeatedly join group-purchasing initiated by a relatively small number of initiators they co-bought with before.

Therefore, we hypothesize that customers' purchasing decisions are not only influenced by their own purchasing preferences but also strongly influenced by who the group initiator is. In the next section, we propose two new models to predict which customers are most likely to join a particular group-purchasing event.

4 Methodology

In this section, we first describe in Section 4.1 how we apply topic modeling to learn user purchasing preferences under the group-purchasing scenario. During the training phase, we compute for each user a mixture topic proportion by combining topic proportions of this user and the initiators with whom this user has co-bought products.

Given a new group-purchasing event, we would like to predict which customers are more likely to join. We propose two predictive models in Section 4.2. One model, which we denote as the product-centric inference model (PCIM), computes the posterior probability that a user would purchase this product given his or her mixture topic proportion. The other model, which we denote the group initiator centric inference model (GICIM), computes the posterior probability that a user would join the group-purchasing event initiated by this initiator given user's or initiator's mixture topic proportion.

4.1 Topic Model for User Purchasing Preference

We use topic modeling to characterize a user's purchasing preference. In particular, we apply LDA to our group-purchasing dataset. In a typical LDA model for text mining [2], a document is a mixture of a number of hidden topics which can be represented by a multinomial distribution, i.e. the topic proportion. A word can belong to one or more hidden topics with different probabilities. Figure 4 shows the graphical model for LDA. LDA is a generative model where each word in a document is generated by two steps: 1) sample a topic from its topic distribution and 2) draw a word from that topic. One can also use Bayesian inference to learn the particular topic proportion of each document.

In our model, we treat a user's purchase history as a document. Each purchased product can be seen as a word in a document. We make an analogy between text documents and purchasing patterns as shown in Table 1. We replace words with each purchased product and a document is one user's purchasing history. Assume that there are U users in our training data. Let $\mathbf{U}$ denote the set of users. Each user $u \in \mathbf{U}$ has a vector of purchased products $\boldsymbol{x}_u = \{x_{un}\}_{n=1}^{N_u}$ where N_u is the number of products that user u purchased.

The generative process of the LDA model for learning a user's purchasing preferences is described as following. Each user u has his or her own topic proportion (i.e., purchasing preference) $\theta_{\mathbf{u}}$ that is sampled from a Dirichlet distribution. Next, for each product x_{un} purchased by user u, a topic z_{un} is firstly chosen from the user's topic proportion θ_u. Then, a product x_{un} is drawn from the multinomial distribution $\phi_{z_{un}}$. To estimate θ_u and $\phi_{z_{un}}$, we use the collapsed Gibbs sampling method [15].

Symbol	Description for Group Purchase History	Description for Text Documents
U	Number of users	Number of documents
K	Number of latent topics	Number of latent topics
N_u	Number of purchased products of user u	Number of words of document u
z_{un}	Latent co-purchasing category of nth product	Latent topic of nth word of document u
x_{un}	nth purchased product of user u	nth word of document u
θ_u	Latent co-purchasing category proportion for user u	Topic proportin for document u
ϕ_k	Multinomial distribution over products for topic k	Mult. distribution over words for topic k
α	Dirichlet prior parameters for all θ_u	Dirichlet prior parameters for all θ_u
β	Dirichlet prior parameters for all ϕ_k	Dirichlet prior parameters for all ϕ_k

Table 1: Latent Dirichlet allocation plate model notation

4.2 Proposed Models for Predicting Group-Deal Customers

A group-purchasing event contains two kinds of critical information: who the group initiator is and what the group-deal product is. Our goal is to predict which customers are more likely to join a specific group-purchasing event. Intuitively, one may think that whether a customer would join a group-purchasing event solely depends on what the group-deal product is. However, from our observations in the dataset, we hypothesize a correlation between a customer's purchasing decision and who the group initiator is. Therefore, we would like to study how these two kinds of group-purchasing information affect the prediction accuracy by asking two questions:

1. What is the likelihood that a customer would join the event given what the *group-deal product* is?
2. What is the likelihood that a customer would join the event given who the *group-initiator* is?

This leads to our two proposed predictive models, the product centric inference model (*PCIM*) and the group initiator centric inference model (*GICIM*).

4.2.1 Product Centric Inference Model (PCIM)

Figure 5(a) shows the graphical structure of PCIM. For each user, we train a PCIM. PCIM computes the posterior probability that a user would purchase a product given his or her mixture topic proportion. Let C denote the *user's own topic proportion*, which we learned from LDA. Suppose that this user has joined group-purchasing events initiated by n group initiators, we use $\{I_i\}_{i=1}^n$ to denote the *learned topic proportions of these initiators*. Our model computes the weighted topic proportions of initiators W by linearly combining $\{I_i\}_{i=1}^n$ with the frequency distribution that the user co-bought products with them.

Intuitively, if a user joins a group-purchasing event initiated by a group initiator, they might share similar interests. Therefore, our model characterizes the user's purchasing preferences by a weighting scheme that combines C and W with a weighting parameter w. We use M to denote such a *mixture topic proportion* which encodes the overall purchasing preferences of the user.

Let P denote the *product random variable*. $\Omega(\mathrm{P}) = \{p_1, ..., p_m\}$, where p_i is the product. From each data record $D_i \in \boldsymbol{D}$, we have a tuple (t_i, p_i, o_i, $\mathbf{m_i}$) and know what group-deal product p_i corresponds to a particular group-purchasing event e_i. Our goal is to compute $Pr(P = p_i)$, the probability that the user would join a group-purchasing event e_i to buy a product p_i.

Given the topic proportion C and $\{I_i\}_{i=1}^n$ corresponding to the user and the weighting parameter w, we are able to compute

$$Pr(P) = \sum_{\mathbf{Y}=\mathbf{X}_p \setminus \{P\}} Pr(P, \mathbf{Y}) \tag{1}$$

where $\mathbf{X}_p = \{P, M, C, W, I_1, ..., I_n\}$; P a is product random variable; M is a mixture topic proportion.

To predict which users are more likely to join a group-purchasing event, we rank $\{\mathcal{P}_{p_i}^{(u_1)}, \ldots, \mathcal{P}_{p_i}^{(u_U)}\}$ in descending order where $\{u_1, \ldots, u_U\}$ denotes the set of users in our dataset and $\mathcal{P}_{p_i}^{(u_j)}$ denotes $Pr(P = p_i)$ of user u_j.

4.2.2 Group Initiator Centric Inference Model (GICIM)

The graphical illustration of GICIM is shown in Figure 5(b). GICIM computes the posterior probability that a user would join a group-purchasing event initiated by a particular initiator given user's or initiator's mixture topic proportion. GICIM and PCIM only differ in their leaf nodes. While PCIM considers only what the group-deal product is, GICIM models our observation that the decision of whether or not a user joins a

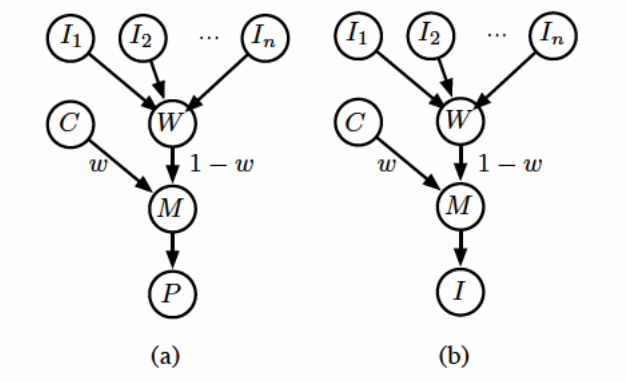


Figure 5: Our proposed models for predicting potential customers given a group-purchasing event. (a) Product centric inference model (PCIM). (b) Group initiator centric inference model (GICIM).

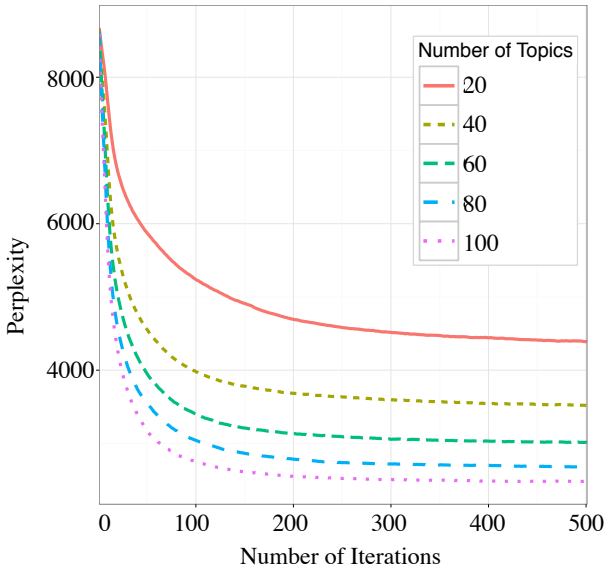


Figure 6: Perplexity as function of collapsed Gibbs sampling iterations for different number of topics used in LDA.

group-purchasing event is strongly influenced by who the group initiator of that event is.

Let I denote the initiator random variable and i_i denote the initiator of the group-purchasing event e_i. Again, from each data record $D_i \in \boldsymbol{D}$, we know who the group initiator is. Instead of evaluating $Pr(P = p_i)$ as in PCIM, we use GICIM to compute

$$Pr(I) = \sum_{\mathbf{Y}=\mathbf{X}_p \setminus \{I\}} Pr(I, \mathbf{Y}) \quad (2)$$

where $\mathbf{X}_p = \{I, M, C, W, I_1, ..., I_n\}$; I is an initiator random variable; M is a mixture topic proportion.

To predict which users are more likely to join a group-purchasing event e_i, we rank $\{\mathcal{P}_{o_i}^{(u_1)}, \ldots, \mathcal{P}_{o_i}^{(u_U)}\}$ in descending order where $\{u_1, \ldots, u_U\}$ denotes the set of users in our dataset and $\mathcal{P}_{o_i}^{(u_j)}$ denotes $Pr(I = o_i)$ of user u_j.

5 Experimental Evaluation

5.1 Data Pre-processing

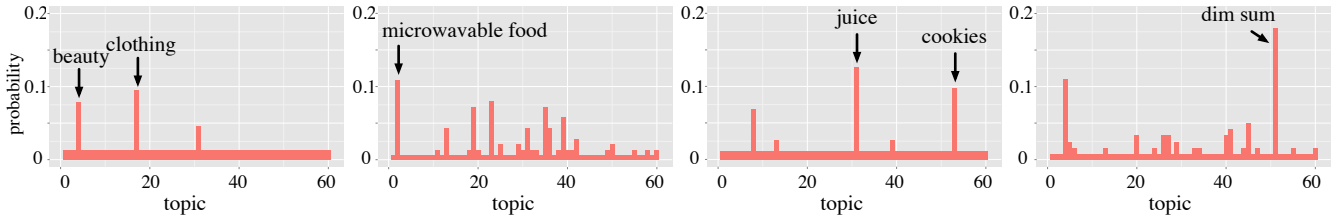
We evaluate the proposed PCIM and GICIM models with the ihergo group-purchasing dataset. In order to capture meaningful user purchasing preferences, we remove users who purchased fewer than 10 products during the pre-processing step. We use ten-fold cross-validation to generate our training and testing datasets.

5.2 LDA Topic Modeling on Group Purchasing Dataset

To measure the performance of LDA for different number of topics (20, 40, 60, 80, 100) in our group-purchasing dataset, we compute the perplexity. It measures how well the model generalizes and predicts new documents [2]. Figure 6 shows that the perplexity decreases as the number of iteration increases and converges within 200 iterations. In addition, as we increase the number of topics, the perplexity decreases. Unless mentioned specifically, all topic proportions used in our experiments are learned with LDA using 100 topics.

Figure 7 shows three example product topics learned by LDA using 100 topics. Each table shows the ten products that are most likely to be bought in that topic. Columns in the table represent the product name, the probability of the product being purchased in that topic, and the ground-truth category of the product, respectively. We see that Topic 1 is about "pasta." It contains a variety of cooked pasta and pasta sauce. Topic 18 and 53 are respectively about "bread and cakes" and "women accessories."

Figure 8 shows the topic proportions of four randomly selected users learned by LDA using 60 topics. We see that different users have distinguishable topic proportions, representing their purchasing preferences. For example, user #3617 purchased many products that are about "beauty" and "clothing" so her or his topic proportion has higher probabilities at topic 4 and topic 17. Similarly, user #39 tends to buy products in the "dim sum" category which can be represented in her or his topic proportion.



(a) User #3617 (b) User #5 (c) User #2618 (d) User #39

Figure 8: Examples illustrating learned topic proportions of four randomly selected users. For user #3617, the two most probable topics are "beauty" and "clothing".

Product	Prob.	Category
Chicken pasta (cream sauce)	0.0197	*Pasta*
Chicken pasta (pesto sauce)	0.0193	*Pasta*
Pork pasta (tomato sauce)	0.0167	*Pasta*
Pork steak	0.0155	*Meat*
Bacon pasta (cream sauce)	0.0153	*Pasta*
Spicy pasta (tomato sauce)	0.0149	*Pasta*
Clam garlic linguine	0.0146	*Pasta*
Tomato sauce pasta	0.0142	*Pasta*
German sausage sauce	0.0132	*Pasta*
Italian pasta (cooked)	0.0129	*Pasta*

(a) Topic 1, "pasta"

Product	Prob.	Category
Ham sandwich	0.0101	*Bread*
Cheese sandwich	0.0089	*Bread*
Milk bar cookie	0.0080	*Cookie*
Cherry chocolate tart	0.0078	*Cake*
Cheese roll	0.0077	*Cake*
Cheese almond tart	0.0074	*Cake*
Taro toast	0.0073	*Bread*
Creme Brulee	0.0073	*Cake*
Raisin toast	0.0071	*Bread*
Wheat ham sandwich	0.0070	*Bread*

(b) Topic 18, "bread and cakes"

Product	Prob.	Category
Knit Hat	0.0169	*Accessory*
Knit Scarf	0.0165	*Accessory*
Legging	0.0133	*Clothing*
Wool scarf	0.0120	*Accessory*
Long Pant	0.0111	*Clothing*
Cotton Socks	0.0099	*Accessory*
Wool Gloves	0.0097	*Accessory*
Facial Masks	0.0090	*Body Care*
Wool socks	0.0088	*Accessory*
Brown knit scarf	0.0081	*Accessory*

(c) Topic 53, "women accessories"

Figure 7: Illustration of product topics learned by LDA using 100 topics. Category is from ground truth.

5.3 Performance of PCIM and GICIM

We use lift charts to measure the effectiveness of PCIM and GICIM for predicting group-purchasing customers. In a lift chart, the x-axis represents the percentage of users sorted by our prediction score and the y-axis represents the cumulative percentage of the ground-truth customers we would predict. For all lift charts shown in this section, we also include two baseline models for comparison. One baseline model is to predict potential customers by *randomly sampling* from the set of users. Therefore, it is a straight line with slope 1.0 on the lift chart. Another baseline model, which we call category frequency, is to predict customers with the most frequent purchase history in a given product category. Specifically, to predict potential customers given a group-deal product category, we rank each customer in descending order of their normalized purchase frequency for the given product category.

Effect of w. We first measure the effect of the weighting parameter w in PCIM, which is shown in Figure 9. The particular w controls how much the user's own topic proportion is used in the mixture topic proportion. For example, $w = 1$ means that only the user's own topic proportion is used as the mixture topic proportion. We see that for all w values, PCIM performs much better than the baseline models between 0% and 25% of the customers predicted. For instance, PCIM is able to reach 50% of the targeted customers while the two baseline models only reach respectively 25% and 40% of the customers.

We also see that with $w = 1$, the curve first rises very fast, then flattens between 25% and 50%. It even performs worse than the baseline models starting at around 60% of the customers predicted; however this is the least interesting part of the curve. The intuition behind this behavior is that with $w = 1$, PCIM is good at predicting customers who have strong purchasing preferences that match the targeted group-deal product. On the other hand, for users without such strong purchasing preferences, the model is not able to per-

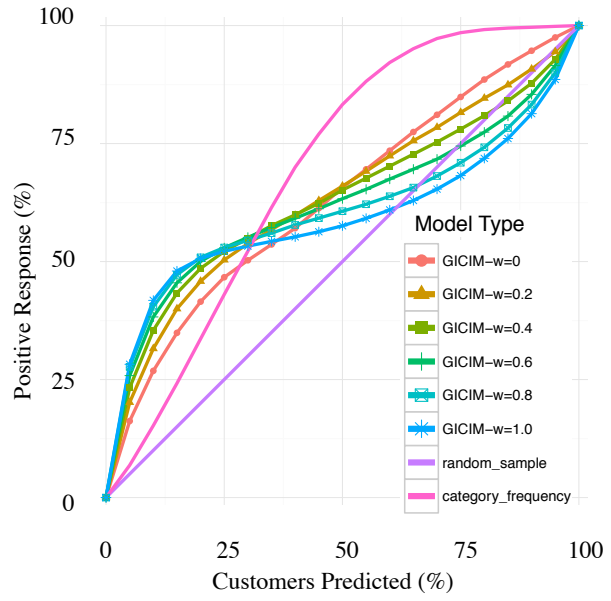


Figure 9: Lift chart of PCIM with different weighting parameter values. With $w = 1$, the model only includes the user's own topic proportion.

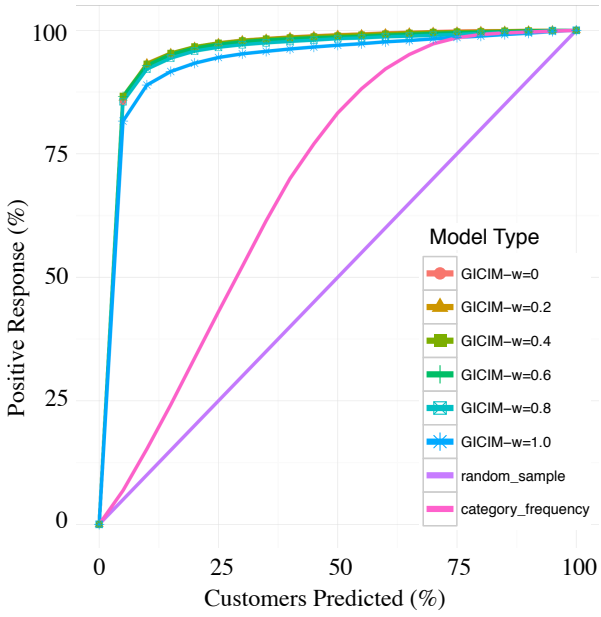
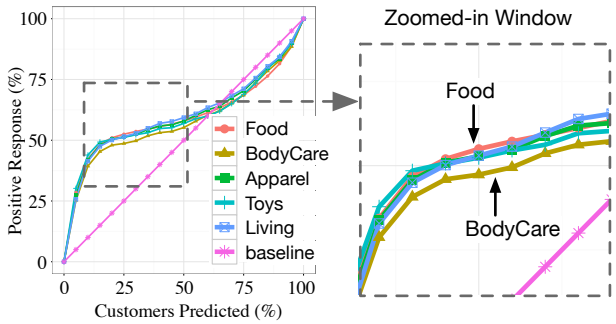
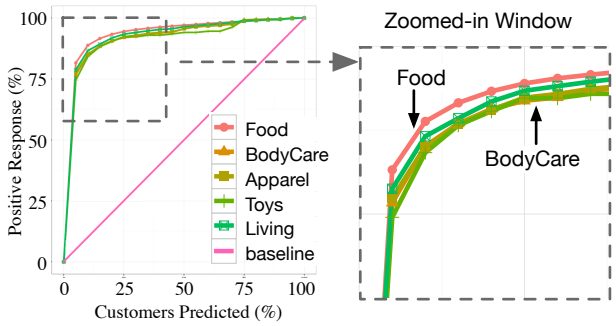


Figure 10: Lift chart of GICIM with different weighting parameter values. With $w = 1$, the model only includes the user's own topic proportion.



(a) PCIM



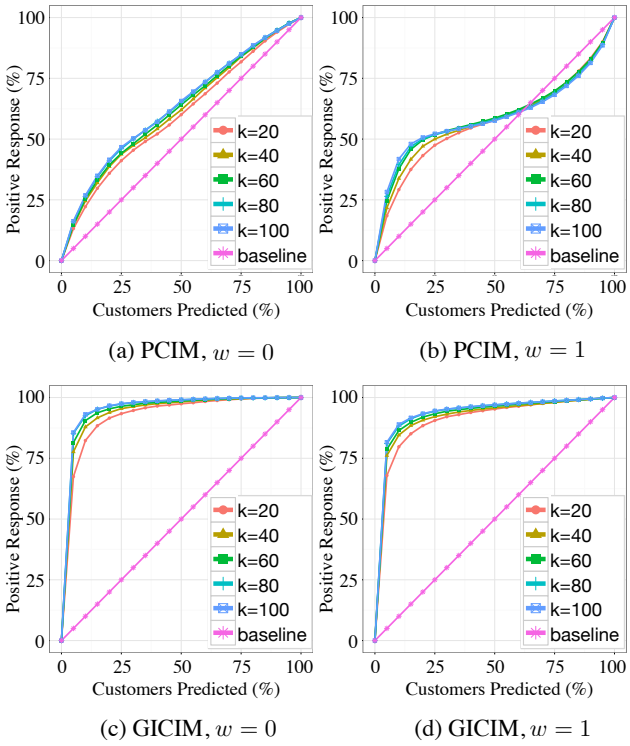
(b) GICIM

Figure 11: Lift charts of PCIM and GICIM over different product categories. The zoomed-in windows on the right show that performance is slightly better on frequently purchased items (*Food*) than on infrequently purchased items (*Body Care*).

form well. In general, by introducing the topic proportions of initiators with whom the user has co-bought products ($w < 1$), PCIM is able to reduce the flattening effect. With $w = 1$, we see that the lift curve is always above the baseline.

Figure 10 shows the effect of w in GICIM. We see that GICIM always performs better than PCIM and the baseline model even for the case where $w = 1$. In particular, for the cases where $w < 0.8$, GICIM achieves 90% positive response with only 10% of the predicted customers. The high prediction success of GICIM can be explained by the fact that whether a user chooses to join a group-purchasing event or not depends on who the group initiator is. We also note that the performance change due to different w values is not as significant as for PCIM.

Performance on different product categories. We next investigate whether frequently purchased items (e.g., drinks and food items) make PCIM and GICIM perform differently. We test on five different categories of group-deal products: *food*, *body care*, *apparel*, *toys*, and *living*. The ground-truth categories are from the dataset.



(a) PCIM, $w = 0$ (b) PCIM, $w = 1$

(c) GICIM, $w = 0$ (d) GICIM, $w = 1$

Figure 12: Lift charts of PCIM and GICIM over different number of topics.

Figure 11 shows the results. We see that, for all product categories tested, GICIM still performs better than PCIM. Moreover, from Figure 11, we see that both models are slightly better at predicting potential customers for the *food* category than for the *body care* category. We hypothesize that this may be related to the fact that purchases in the *food* category are more frequent and predictable compared to purchases in the *body care* category. A customers may buy one or more products in the *food* category repeatedly, while the same does not appear to be the case for all products in the *body care* category. For example, once someone has purchased a sunscreen spray (a product in *body care*), they are probably unlikely to buy it again, at least for the time span that our dataset covers. However, note that the differences between categories in Figure 11 are small, and developing a better understanding of them is an area of future research.

Effect of different number of topics. We ran PCIM and GICIM on different number of topics used in LDA. Results are given in Figure 12. We find that increasing the number of topics increases prediction accuracy for both models. This agrees with the above perplexity analysis that higher number of topics results in better performance.

6 Conclusion

In this paper, we study group-purchasing patterns with social information. We analyze a real-world group-purchasing dataset (5,602 users, 26,619 products, and 13,609 events) from ihergo.com. To the best of our knowledge, we are the first to analyze the group-purchasing scenario where each group-purchasing event is started by an initiator. Under this kind of social group-purchasing framework, each user builds up social ties with a set of group initiators. Our analysis of the dataset shows that a user usually joins group-purchasing events initiated by a certain and relatively small number of initiators. That is, if a user has co-bought a group-deal product with a group initiator, he or she is more likely to join a group-purchasing event started by that initiator again.

We develop two models to predict which users are most likely to join a group-purchasing event. Experimental results show that by including the weighted topic proportions of the initiators, we achieve higher prediction accuracy. We also find that whether a user decides to join a group-purchasing event is strongly influenced by who the group initiator of that event is.

Our model can be further improved in several ways. First, we can use Labeled LDA [16] by exploiting the ground-truth category of the products or user profile from the dataset. Second, we can incorporate other information such as the geographical and demographic information of users, and the seasonality of products in a more complex topic model. We are also interested in investigating the model to deal with *cold start*, where a new user or group-deal product is added to the system.

Acknowledgments

We want to thank Kuo-Chun Chang from ihergo.com for his cooperation and the use of ihergo dataset. This material is based, in part, upon work supported by NSF award CCF0937044 to Ole Mengshoel.

References

[1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. on Knowl. and Data Eng.*, 17(6), June 2005.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3, Mar. 2003.

[3] J. W. Byers, M. Mitzenmacher, M. Potamias, and G. Zervas. A month in the life of Groupon. *CoRR*, 2011.

[4] J. W. Byers, M. Mitzenmacher, and G. Zervas. Daily deals: Prediction, social diffusion, and reputational ramifications. *CoRR*, abs/1109.1530, 2011.

[5] Y. Cha and J. Cho. Social-network analysis using topic models. SIGIR '12, 2012.

[6] B. Edelman, S. Jaffe, and S. D. Kominers. To groupon or not to groupon: The profitability of deep discounts. Harvard business school working papers, Harvard Business School, Dec. 2010.

[7] S. Guo, M. Wang, and J. Leskovec. The role of social networks in online shopping: Information passing, price of trust, and consumer choice. *CoRR*, abs/1104.0942, 2011.

[8] T. Hofmann. Collaborative filtering via gaussian probabilistic latent semantic analysis. SIGIR '03, 2003.

[9] T. Iwata, S. Watanabe, T. Yamada, and N. Ueda. Topic tracking model for analyzing consumer purchase behavior. In *IJCAI*, 2009.

[10] T. Iwata, T. Yamada, and N. Ueda. Modeling social annotation data with content relevance using a topic model. In *In NIPS*, 2009.

[11] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web*, 1(1), May 2007.

[12] T. Lu and C. E. Boutilier. Matching models for preference-sensitive group purchasing. EC '12. ACM, 2012.

[13] H. Ma, I. King, and M. R. Lyu. Learning to recommend with social trust ensemble. SIGIR '09. ACM, 2009.

[14] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King. Recommender systems with social regularization. WSDM '11. ACM, 2011.

[15] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling. Fast collapsed gibbs sampling for latent dirichlet allocation. KDD '08. ACM, 2008.

[16] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. EMNLP '09. Association for Computational Linguistics, 2009.

[17] H. M. Wallach. Topic modeling: beyond bag-of-words. ICML '06. ACM, 2006.

[18] X. Yang, Y. Guo, and Y. Liu. Bayesian-inference based recommendation in online social networks. *IEEE Transactions on Parallel and Distributed Systems*, 99(PrePrints), 2012.

[19] M. Ye, C. Wang, C. Aperjis, B. A. Huberman, and T. Sandholm. Collective attention and the dynamics of group deals. *CoRR*, abs/1107.4588, 2011.

[20] L. Yu, R. Pan, and Z. Li. Adaptive social similarities for recommender systems. RecSys '11. ACM, 2011.