GroupBox: A generative model for group recommendation

Cheng Zhang
KTH Royal Institute of Technology
Stockholm, Sweden
chengz@kth.se

Mike Gartrell
Microsoft R&D
Herzliya, Israel
mike.gartrell@acm.org

Thomas P. Minka, Yordan Zaykov, John Guiver
Microsoft Research
Cambridge, United Kingdom
{minka, yordanz, joguiver}@microsoft.com

Technical Report
MSR-TR–2015-61

GroupBox: A generative model for group recommendation

Cheng Zhang *
KTH Royal Institute of Technology
Stockholm, Sweden
chengz@kth.se

Mike Gartrell
Microsoft R&D
Herzliya, Israel
mike.gartrell@acm.org

Thomas P. Minka, Yordan Zaykov, John Guiver
Microsoft Research
Cambridge, United Kingdom
{minka, yordanz, joguiver}@microsoft.com

ABSTRACT

In this paper, we present a principled probabilistic framework – GroupBox – for making recommendations to groups. GroupBox is able to model user influence within a group, the suitability of an item to a group context, and the differences in user preference between individual and group contexts. Efficient scalable inference algorithms are used for GroupBox, which makes it applicable to large-scale datasets. We run experiments on a large-scale TV viewing dataset collected by Nielsen and show how the model can be used to understand both context and influence. The experimental results on the large scale real data provide a deep understanding of the individual behaviours in group context.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

group recommendation, hybrid recommender, cold-start, TV program viewing patterns

1. INTRODUCTION

Recommendation systems are an important class of applications widely used in practice, for example to recommend videos on YouTube/Netflix, to recommend Music on Spotify/Pandora, or to recommend products on Amazon/Alibaba. Most of the current recommendation systems are designed for individual users only. However, group activities are common; for example, family or friends watching a movie together, or friends having a party with background music. As explained later, US TV viewing data collected by Nielsen shows that around 25% of all viewing events are group viewings. Recommendation systems for groups are not as prevalent as those for individuals, perhaps due to the lack of data (which ideally would include group ratings and group make-up) and due to the higher complexity of the problem. In this paper, we focus on developing a recommendation system that is applicable to both individuals and groups.

Recommendation systems for individuals have a long history and can be divided into content-based methods, collaborative filtering methods, and hybrid methods that combine the previous two. Many different such models have been developed. Generative models, in general, have an advantage in terms of dealing with missing data and providing a better understanding of the latent structure in the data. In this work, we will use a hybrid method which can utilize metadata about individuals, items, and groups. Our generative model allows a deep understanding of user and item properties. Our work extends MatchBox [17], which is a hybrid generative recommendation framework designed for personalized individual recommendation, and we refer to this extension as GroupBox.

Group decisions are different from individual choices in ways that are not explained by simply combining recommendations for the individuals in the group. A person may desire to watch a different TV program when they are in a group than when alone, and this may differ depending on the group composition, such as a group of friends versus a group of siblings and their parents. We refer to this as contextual item preference. Even without contextual preference, group members may have different influence in groups [3] depending on their roles in the group, for example, an elder brother and a young sister may have different influence on the decision of which TV program to watch. Modelling this influence is not only useful for making better group recommendations, but also for understanding the social behaviour of people in groups.

Our key contributions in this paper are as follows:

1. We propose a generative model for large scale group recommendation, which can be applied to both group users and individual users. (Section 4.2)
   (a) The model can distinguish between a individual's preferences in a group context versus a solitary context.
   (b) The model can distinguish between an individual's preferences in one group or group type versus another.
   (c) The model can determine the influence of a user within a group.
   (d) The model can distinguish between the popularity of an item in a group context versus a solitary context.
2. We evaluate our model on a large-scale dataset collected by Nielsen and show how the model can be used to provide better insights into real-life group viewing patterns. (Section 6.2)

2. PROBLEM STATEMENT

We address the problem of group recommendation, which seeks to predict the rating or preference that a group of users would give to an item. In this domain we distinguish between an individual instance, which is an item viewed and rated by an individual user, and a group instance, where a collection of users has collectively viewed an item and provided a single group rating for this item. In

*This work has been done during her internship at MSRC.

group recommendation, we learn from these individual and group instances to predict group ratings.

In this paper, we use TV viewing data to get individual and group ratings. The data consists of users, households, TV programs, and TV program views. These entities have associated metadata, such as age and gender for users, and distributor and genre for programs. We define an item in our system as composed of a program name, program start time, program distributor, and genre. As a result of this definition, different episodes of a TV series are treated as a single item, since these episodes are commonly broadcast on the same channel (program distributor) at the same time. For each program view, we have the start time and duration of the view, along with the user and household who viewed the program. From this data we construct group program views, where a group program view is defined as an item that is viewed by a fixed set of more than one user in the same household for a specified program start time and date. We define an individual program view as an item that is viewed by only one user in a household for a specified program start time and date. Each view is weighted by the fraction of the program viewed by the individual or group.

Our model requires "positive" and "negative" data for training. All individual and group views are considered positive data. To construct a negative view instance, for each positive view instance we begin by examining those items whose broadcast overlaps with that view. We sample one negative view instance from this set of overlapping items that have not been viewed according to the relative viewing popularity of the items in this set; see Section 6.2.1 for further details on this negative sampling scheme. Given the above definitions of positive and negative data, we seek to predict group preference by computing the probability that a group will view a particular item.

The *cold-start* problem is common in recommender systems. In our domain, cold-start users are defined as users who have not yet viewed any items. Cold-start items are items that have not been viewed by any users or groups. Finally, cold-start groups are defined as groups who have not viewed any items together. Note that cold-start groups may or may not contain cold-start users.

3. RELATED WORK

Systems that make recommendations for groups have been around since the late nineties. The opportunity to apply such systems to a variety of areas (such as tourism, dining, movies, television, video games, and music selection) is becoming increasingly practical in today's world of portable online identities and instant social communication. Such systems typically need to know the individuals in a group, but the group itself may be previously known or ad hoc. [9] gives a summary of the state of the art in 2007 and categorizes group recommendation systems into three categories: (1) aggregation of individual sets of recommendations, (2) aggregation of predictions of individual ratings, and (3) group modelling. A fourth strategy is the aggregation of user profiles into a group profile to create a new 'individual' profile.

The second category in particular is widely used and makes use of a variety of aggregation functions including *average satisfaction*, *least misery*, and *maximum satisfaction* [9]. *Average satisfaction* assigns equal importance to each group member and is used in [6, 21, 20]. Different user weights, dissimilarity among group members, social connections, and personality profiles are also used in aggregation models [5, 1, 8, 14]. [13] does a systematic evaluation of all combinations of several common individual recommender models with several common aggregation methods. Their conclusion is that the best-performing aggregation method depends on the choice of individual recommender; moreover, different combinations favour different goals such as accuracy, coverage, diversity, and novelty.

Due to the paucity of realistic data sets with member identities and group labels, many of the aggregation-based models are run on individual recommender data sets by synthetically generating random groups. However in such studies the accuracy evaluation metrics are somewhat contrived because they, like the aggregation methods themselves, are formulated in terms of individual preferences [4]. In particular such evaluations do not take into account the make-up of the group or the fact that some items are more suited to group consumption than others.

However, large-scale group preference datasets with group labels are beginning to emerge. The 2011 Challenge on Context-Aware Movie Recommendation (CAMRa 2011) used a dataset consisting of about 170,000 users, 24,000 movies, and 4.4 million ratings [16]; in this dataset, a group is synonymous with a household, and members of a household are known, but there are group ratings from only 290 households. In our work we assume that there are explicit or implicit labels for group ratings. In practice, and especially with groups, only positive implicit labels may be available (corresponding to user consumption of the item by the group); in this case there are strategies to generate negative labels [15, 12]. In particular, we evaluate our method using a dataset [7] containing hundreds of thousands of implicit group preferences, along with substantial metadata for individuals, groups, and items.

The individual recommender models that participate in the aggregation approaches to group recommendation are too numerous to review here, but at the highest level can be split into the well-known approaches of collaborative filtering (CF) versus content-based (CB). Hybrid models such as Matchbox [17] or topic models [19] give the best of both worlds combining the accuracy of CF with the cold-start capabilities of CB models. These concepts extend to the construction of group models (the third category of models in [9]) though it is important to note that there are additional considerations in a group setting some of which were discussed in section 2.

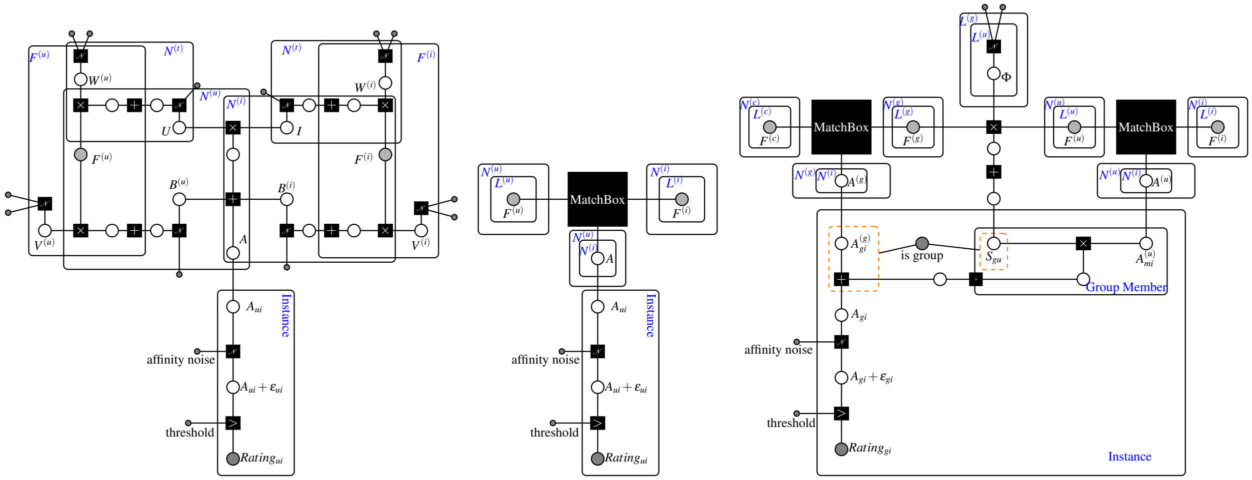
There are various aspects that one want to model in a group context. For example, different members will have different levels of assertiveness or acquiescence; individuals will have different preferences in a group context; and certain items may be more suited to a particular group context. [2] requires a history of shared consumption and uses this to determine influence within the group as well as the tendency of the group to watch certain genres. [22] uses a probabilistic generative model in which each individual has a preference distribution over all items. Here is no persistent notion of group; instead each group instance is assumed to independently draw a topic and set of users, and the user preferences aggregate to choose the item. Instances with 1 person are not distinguished, so this model cannot account for someone having a different preference in a group context. Our approach, inspired by Matchbox, is a scalable generative probabilistic model, combining CF and CB approaches, that has explicit variables for persistent groups, user influence, and preference changes due to group context.

4. MODEL

In this section, the Matchbox model for individual recommendation, on which our group recommender is based, is revisited first. Then we will present our GroupBox model which can be used for both individual recommendations and group recommendations. Factor graph representations of the models are shown in Figure 1 with an explanation of the symbols used in Table 1.

4.1 Matchbox for individual recommendation

MatchBox [17] is a generative model designed for personalized



(a) MatchBox for individual recommendation (b) Simplified MatchBox representation (c) GroupBox for group recommendation

Figure 1: Factor Graphs.

	Variable	Symbol
Observed	User, Item, Group, ItemContext Feature	$F^{(u)}$,$F^{(i)}$,$F^{(g)}$,$F^{(c)}$
Latent	User Item, Group, Item Context Traits	U,I ,G ,C
	User, Item, Group, Item Context trait feature weight	$W^{(u)}$, $W^{(i)}$, $W^{(g)}$,$W^{(c)}$
	User, Item, Group, Item Context Bias	$B^{(u)}$,$B^{(i)}$,$B^{(g)}$,$B^{(c)}$
	User, Item, Group, Item Context Bias feature weight	$V^{(u)}$,$V^{(i)}$,$V^{(g)}$,$V^{(c)}$
	User Influence	Φ
	Individual, Group, Overall Affinity	$A^{(u)}$,$A^{(g)}$,A
Other	number of Traits for individual/group MatchBox	$N^{(t)}$, $N^{(t')}$
	number of users, items, groups, contextual items	$N^{(u)}$, $N^{(i)}$, $N^{(g)}$, $N^{(c)}$
	size of the user, item, group, contextual item feature vectors	$L^{(u)}$, $L^{(i)}$,$L^{(g)}$,$L^{(c)}$

Table 1: List of Symbols

recommendation for individual users. It is a hybrid system which combines content based modelling and collaborative filtering. It has achieved good performance and is able to handle cold-start situations. MatchBox is a matrix factorization based method. The axes of the low dimensional latent space are called traits which represent latent properties of the data. The model assumes that the affinity A, a $N^{(u)} \times N^{(i)}$ matrix where each element represents a person's opinion about an item, is generated as:

$$A_{ui} = \sum_t U_{ut} I_{ti} + B_u^{(u)} + B_i^{(i)}, \tag{1}$$

where user trait U is a $N^{(u)} \times N^{(t)}$ matrix, item trait I is a $N^{(t)} \times N^{(i)}$ matrix, $B^{(u)}$ is an $N^{(u)}$ sized vector and $B^{(i)}$ is an $N^{(i)}$ sized vector. All elements in these matrices are generated from independent Gaussian distributions, whose variance is a hyperparameter and whose mean is computed from features. For example:

$$p(U) = \prod_{u=1}^{N^{(u)}} \prod_{t=1}^{N^{(t)}} \mathcal{N}(m_{ut}, \sigma^2) \tag{2}$$

$$m_{ut} = F_u^{(u)T} W_t^{(u)} \tag{3}$$

where $F_u^{(u)}$ is a vector of features and $W_t^{(u)}$ is a vector of weights. The features are observed and the weights need to be inferred, and we assume that each element of the user feature weight matrix $W^{(u)}$ is generated independently from a Gaussian distribution whose mean and variance are given by hyperparameters. The same approach is used to generate the item trait matrix I, user bias B^u and item bias $B^{(i)}$. Observations are commonly noisy and thus standard Gaussian noise ε is added to the generated affinity A before thresholding to get the rating.[1]

$$\text{Rating}_{ui} = 1(A_{ui} + \varepsilon_{ui} > 0). \tag{4}$$

4.2 GroupBox for group recommendation

We propose the GroupBox recommendation system for groups. An item is called a contextual item if it is watched in groups and the corresponding feature is called item context feature. The item features are observed and the item context feature is the same as the item feature for the same item. In this model, we use a MatchBox sub-model for individual preferences with $N^{(t)}$ traits and another MatchBox sub-model with $N^{(t')}$ traits to model group preferences, where each group is treated as a whole unit. These two MatchBox sub-models have independent trait spaces; in this way GroupBox is able to model the difference between group activities and individual activities. Group bias and item context bias are both modelled in the group MatchBox part. The group bias is modelled for the group as a single unit, and the additional item preference in group scenarios are modelled by item context bias. The model also includes a user influence variable Φ which encodes how much a certain type of user contributes to the decision of a specific type of group. For an individual instance, (1) computes the affinity. For a group instance, the overall affinity is the sum of the group-specific affinity and the weighted summation of the group member individual affinities. The overall affinity for a group g with item i is:

$$A_{gi} = \sum_{t'} G_{gt'} C_{t'i} + B_g^{(g)} + B_i^{(c)} + \sum_{m \in g} S_{gm} A_{mi}^{(u)} \tag{5}$$

where the weight S is computed by multiplying user feature $F^{(u)}$, group feature $F^{(g)}$ and user influence Φ:

$$S_{gm} = F_g^{(g)T} \Phi F_m^{(u)}, \tag{6}$$

[1] MatchBox can be applied to different types of rating data. In this paper, we concentrate on the binary rating situation with threshold 0 due to the nature of the Nielsen data. However, both MatchBox and GroupBox can be easily applied to other types of ratings.

We use a Gaussian prior for each component of the user influence matrix Φ, where the mean and variance are hyperparameters.

5. INFERENCE

In [17], inference was done using a hybrid of Expectation Propagation [11] and Variational Message Passing [18]. We take the same approach here, except we use Infer.Net [10] to generate the schedule and iterate to convergence.

Symmetry breaking is an important issue that was not addressed in previous work. The model has symmetries in the sense that we can transform the parameters into an equivalent set that have identical likelihood. In MatchBox, we can multiply the user traits by a $N^{(t)} \times N^{(t)}$ matrix while multiplying the item traits by the inverse matrix without changing the likelihood. We can also shift the traits and apply an opposite shift to the biases, giving another $2N^{(t)}$ symmetries. Finally, we can shift the user biases while applying an opposite shift to the item biases, giving another symmetry. In GroupBox, the number of symmetries more than double. There are $(N^{(t)} + N^{(t')})^2$ dimensions of symmetry between the four trait matrices, $2N^{(t)} + 2N^{(t')}$ dimensions of symmetry between traits and biases, and 3 shift symmetries between the four biases. These symmetries are evident when trying to recover parameters from synthetic data, since the recovered parameters will generally not match the ones used to generate. If these symmetries are not broken, inference can give poor results, evidenced by excessively large posterior variances due to spanning multiple symmetrical solutions. To prevent this, we break all of the above symmetries by fixing a subset of the traits and biases, according to the number of symmetry dimensions listed above.

6. EXPERIMENT

Experiments are conducted in two stages. Firstly, synthetic data are generated to validate our model. Secondly, large scale real life viewing data are used to explore group viewing patterns. We will present the first experiment only briefly and mainly focus on the second stage of the experiment.

6.1 Synthetic Data

The aim of this experiment is to evaluate the inference performance of the model since the ground-truth is unknown for real-world data. The features are sampled randomly using fixed dimensions. The synthetic viewing instances are sampled using the GroupBox from fixed prior. Then the experiment is conducted using the generated ratings and features to estimate all the latent variables in the model. For the hyper-parameter setting, if we set the trait or bias prior variance much bigger than the corresponding feature weight variance, the data can be explained regardless of the feature weights, hence the feature weights can not be recovered. On the other hand, if we set the prior variance on the feature weights much bigger than the prior variance on the traits and bias, the model can't adapt to individual behavior. In the experiment, the trait prior variances are all set to $\sigma^2 = \frac{1.0}{\sqrt{N^{(t)}}}$ and the bias prior variances are all set to $\nu^2 = 0.05$; the corresponding feature weight variances are set to $10 * \sigma^2$ and $10 * \nu^2$; the affinity noise prior variance is set to 2% of the affinity variance computed by the priors. We are able to recover all the latent parameters with this parameter setting. The parameter recovering results are plotted and shown in the supplement material

6.2 Nielsen Dataset experiment

In this experiment, we use the TV viewing data collected by Nielsen as used by [7]. Using this real world dataset, we explore large-scale viewing patterns by both individuals and groups using our model, GroupBox.

$F_u^{(u)}$	feature Description	$F_u^{(u)}$	feature Description
1	Female 0-11 years old	7	Male 0-11 years old
2	Female 12-25 years old	8	Male 12-25 years old
3	Female 26-38 years old	9	Male 26-38 years old
4	Female 39-50 years old	10	Male 39-50 years old
5	Female 51-62 years old	11	Male 51-62 years old
6	Female 63-99 years old	12	Male 63-99 years old

Table 3: User Feature List

Child	0 - 11 years old
Youth	12 -25 yeas old
Adult	26 -62 years old
Senior	63-99 years old

Table 4: Group age bins

6.2.1 Nielsen Data Set

In this dataset, 4,331,851 viewing instances were recorded in the U.S. in June, 2012. In the experiment, programs which have no overlaps with other programs are removed, since watching such a program does not show any preference if it is the only one available at the time; programs whose broadcast cuts across two days are removed for simplicity; viewing instances with group size larger than 4, corresponding to only 0.15% of the data, are removed to simplify the model. The remaining 2,862,632 viewing instances are used. These viewing instances are all considered as positive with instance weight defined as the fraction of the program being watched. Instance weights are implemented by raising the likelihood of the instance to a power. Items are defined by program name, program starting time, program distributor and genre. In this way, different episodes of the same TV series are treated as the same item since they are commonly broadcast in the same channel starting at the same time; but different seasons of the same TV series are treated as different items since they are commonly broadcast either on different channels or starting at different times.

Negative Data Generation.

All the viewed instances are treated as positive data. A program which is not being watched might be due to either the user(s) disliking it or to them not being aware of the program being aired. We sample the negative instances in the same way as [12, 7]. For each positive viewing instance, we sample one negative viewing instance from all the overlapped programs not being watched. A multinomial distribution is built based on how many times these programs have been watched through the whole dataset. The negative viewing instance is sampled according to this multinomial distribution. Together with the negative data, there are 5,725,264 viewing instances, 8043 items, 71208 users and 81712 groups.

Features Design.

The item and item context feature are defined using the item genre. There are 34 different genres in the dataset. The item feature entries are listed in Table 2. The user features are defined using all combinations of age (6 bins) and gender, which results in 12-dimensional sparse vectors. The user feature entries are listed in Table 3. The group features are defined as group size (2,3,4), gender (only male, only female and mixed) and age (4 age bins + 7 interacted age bins) fully interacted, which results in 96-dimensional sparse vectors. The groups age bins are listed in Table 4.The full list of these features are available in the supplement material.

	F 0–11	F 12–25	F 26–38	F 39–50	F 51–62	F 63–99
Adult			0.298	0.339	0.363	
Child & Youth	0.341	0.659				
Child & Adult	0.158		0.279	0.291	0.273	
Child & Senior	0.394					0.606
Youth & Adult		0.215	0.229	0.258	0.298	
Youth & Senior		0.391				0.609
Adult & Senior			0.177	0.233	0.236	0.354

(a) only female groups

	M 0–11	M 12–25	M 26–38	M 39–50	M 51–62	M 63–99
Adult(26–62)			0.293	0.327	0.380	
Child & Youth	0.390	0.610				
Child & Adult	0.160		0.284	0.274	0.281	
Child & Senior	0.419					0.581
Youth & Adult		0.206	0.216	0.268	0.310	
Youth & Senior		0.382				0.618
Adult & Senior			0.183	0.323	0.264	0.231

(b) only male groups

	F 0–11	F 12–25	F 26–38	F 39–50	F 51–62	F 63–99	M 0–11	M 12–25	M 26–38	M 39–50	M 51–62	M 63–99
Child(1–11)	0.413						0.587					
Youth(12–25)		0.495						0.505				
Adult(26–62)			0.133	0.145	0.169				0.166	0.188	0.199	
Senior(62–99)						0.473						0.527
Child & Youth	0.200	0.287					0.215	0.298				
Child & Adult	0.060		0.129	0.140	0.136		0.081		0.165	0.124	0.165	
Child & Senior	0.150					0.292	0.253					0.304
Youth & Adult		0.091	0.118	0.137	0.140			0.110	0.125	0.135	0.145	
Youth & Senior		0.159				0.348		0.187				0.306
Adult & Senior			0.053	0.100	0.119	0.145			0.136	0.141	0.140	0.165

(c) mixed gender groups

Figure 2: Learned user influence of different type of size 2 groups. Each row is normalized for easy comparison between rows. We show only groups that are composed of different types of users since the same type of user in the same group has the same user influence.

$F_i^{(i)}$	Program type name	Code	$F_i^{(i)}$	Program type name	Code	$F_i^{(i)}$	Program type name	Code	$F_i^{(i)}$	Program type name	Code
1	participation variety	PV	10	daytime drama	DD	19	general variety	GV	28	child multi-weekly	C
2	feature film	FF	11	adventure	A	20	political	P	29	sports news	SN
3	situation comedy	CS	12	sports commentary	SC	21	general documentary	DO	30	quiz panel	QP
4	popular music	PC	13	sports event	SE	22	conversations, colloquies	CC	31	western drama	EW
5	news	N	14	concert music	CM	23	unclassified	U	32	general drama	GD
6	award ceremonies	AC	15	sports anthology	SA	24	science fiction	SF	33	sports news	PD
7	official police	OP	16	devotional	D	25	instruction, advice	IA	34	audience participation	AP
8	evening animation	EA	17	quiz give away	QG	26	popular music	DN			
9	format varies	FV	18	suspense/mystery	SM	27	comedy variety	CV			

Table 2: Item Feature List

6.2.2 *Large-scale viewing pattern analysis using GroupBox*

GroupBox is designed to both understand the group viewing pattern and to recommend programs to users. We firstly use the whole dataset to train GroupBox and explore the learned posteriors to understand the group viewing patterns. The number of traits are set to $N^{(t)} = N^{(t')} = 5$, all other parameters are set in the same way as the synthetic experiment with the same motivation.

Large-scale viewing patterns.

How different types of users contribute in different types of groups is learned through user influence variable Φ using GroupBox. Figure 2 shows the posterior of user influence[2] for the Nielsen data. Many interesting observations can be made from in these plots. For example, youth is in general more dominant in a mixed child and youth group regardless of gender composition; a female youth and a male youth influence the group decision equally but a male adult influences the group decision more than a female adult; seniors are in general more dominant in all different type of groups.

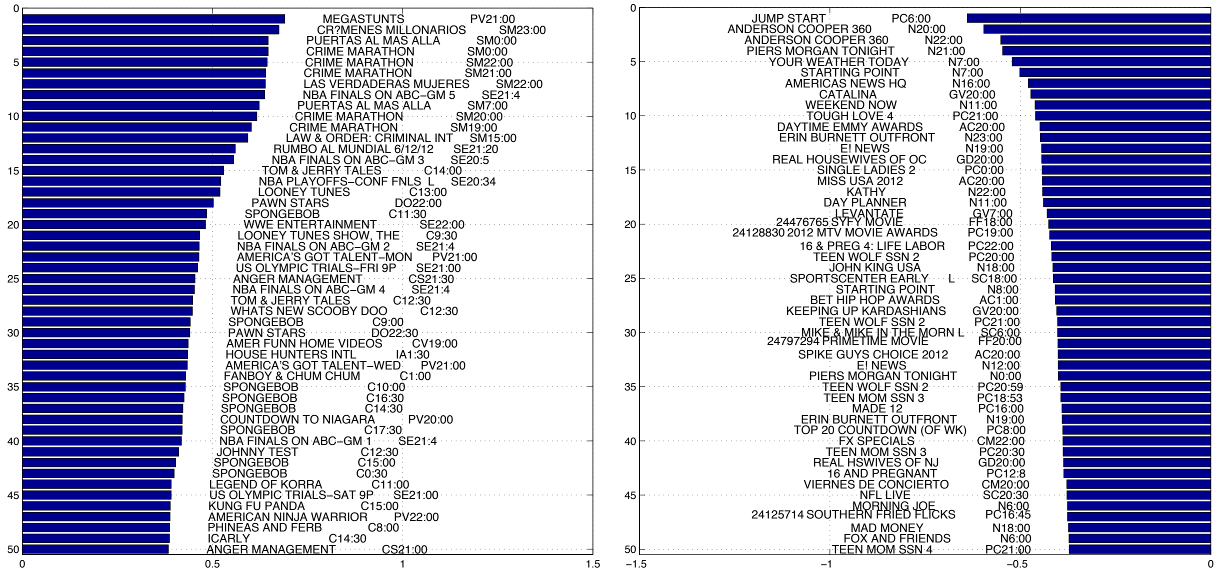
Item bias reflects the popularity of an item among individuals and item context bias reflect the popularity of the item among groups. An item with high item context bias indicates that this item is more popular for group viewing events compared to individual viewing events. Figure 3 shows the histogram of the occurrence of different type of items using the top 500 items with the highest item context bias and top 500 items with lowest item context bias . Results shows that items with high item context biases are commonly family friendly e.g. child multi-weekly (C) [3] and items with low item context biases are commonly popular music (PC) . Figure 4 shows the top 50 item instances with the highest and lowest item context bias. We can see that SpongeBob is one of the most popular items for viewing in groups. Among these 50 items with highest item context bias, 8 of the have the program name SpongeBob. Figure 5 shows the number of positive and negative viewing instances for groups and for individuals for SpongeBob broadcast at different times. The data show that SpongeBob is preferred more by groups than the individuals, which is consistent with the item context bias learned by GroupBox.

Traits should reflect the latent character of the data, and those that we learn from the Nielsen data clearly reflect some meaningful structure. We visualize the learned traits by plotting the first 2 principle components computed using PCA. Item traits and item context traits are visualized in Figure 6. The first two components can only explain around half of the variations. However, we can see that different type of programs are occupying different part of the trait space. For example, in Figure 6 (a) We can clearly see that the upper left part of the item traits plot and the upper right part of the item context traits plot, where news(N) data are located, are more serious than the opposite side of the trait space, where general drama(GD) data are located. Even though we assume that program types are independent, Sport Event (SE) programs and Sport Commentary (SC) locate very close together in the trait plot, which means the traits are capturing natural properties of the data. The structure of the item traits and item context traits are different, which indicates that the group trait space is needed for group recommendation.

Next, we analyze the intra-group behaviours by regret analysis. Regret is a novel visualization that extends the Power Balance Map

[2]Groups with size larger than 2 are shown in the supplement.

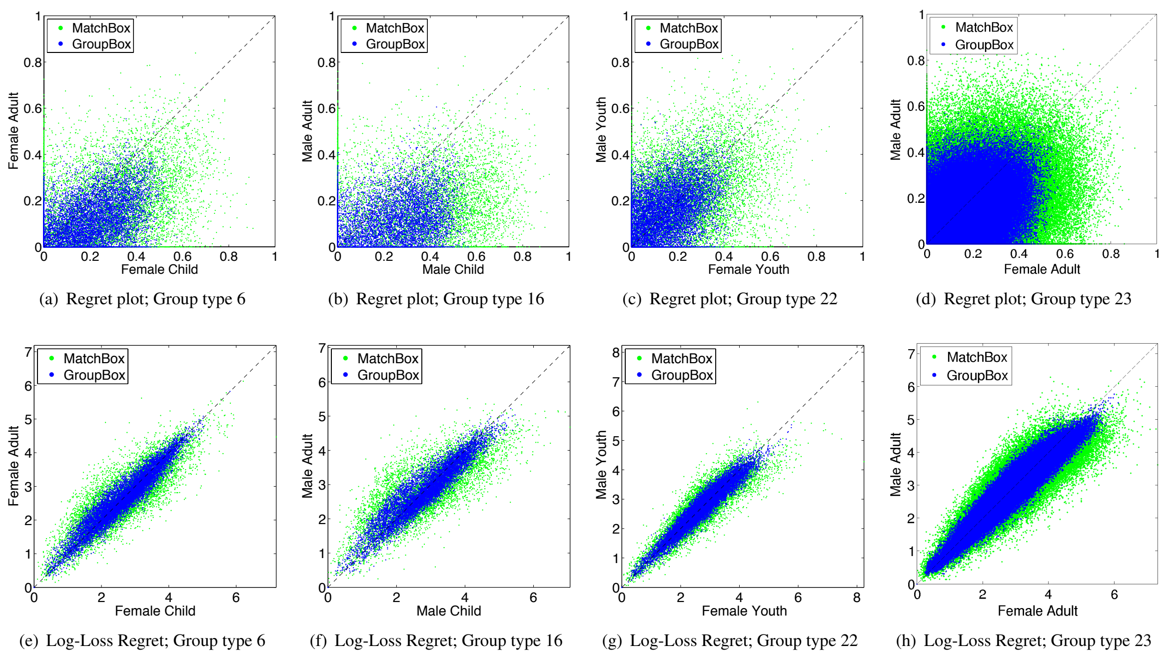
[3]For the rare item types, for example, Suspense/Mystery (SM), which only contains 10 items in total among 8043 items, even they have have high occurrence ratio, we do not consider them significant to present the preference level of a type.



(a) Top 50 items with highest item context bias

(b) Top 50 items with lowest item context bias

Figure 4: Item context bias. Corresponding items are labeled with the item name, program type and starting time. Program name, program type and broadcast stating time are labeled since they are part of the item definition. With this definition, items with the same name but broadcast with different starting time are treated as different items.



(a) Regret plot; Group type 6

(b) Regret plot; Group type 16

(c) Regret plot; Group type 22

(d) Regret plot; Group type 23

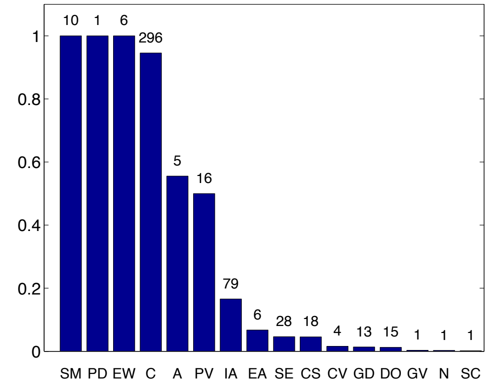
(e) Log-Loss Regret; Group type 6

(f) Log-Loss Regret; Group type 16

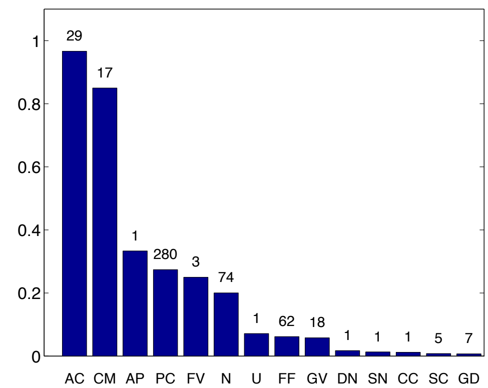
(g) Log-Loss Regret; Group type 22

(h) Log-Loss Regret; Group type 23

Figure 7: Regret plots for different group types. Group type 6: size 2 group with one female child and one female adult; Group type 16: size 2 group with one male child and one male adult; Group type 22: size 2 group with one female youth and one male youth; Group type 23: size 2 group with one female adult and one male adult. In (a) (b) (c) (d), regrets are computed using equation 7. In (e) (f) (g) (h), regrets are computed using equation 8 .



(a) Histogram of 500 items with highest item context bias



(b) Histogram of 500 items with lowest item context bias

Figure 3: Histogram of the items over program types. The x-axis shows the programs types and the y axis shows the rate of occurrence of different item types of the top 500 items with highest and lowest item context bias. The histogram is normalised by the total number of items in each class and the actual count in marked over each bar. Only the non-zero accumulators are plotted in descending order of the occurrence rate.

of [2]. Let $I = \{i_1, i_2 \dots\}$ be all programs available at a given time and i_v is the program viewed by a group at this time. We define the regret for each group member watching the program as:

$$r = max_{i \in I} p(i) - p(i_v) \tag{7}$$

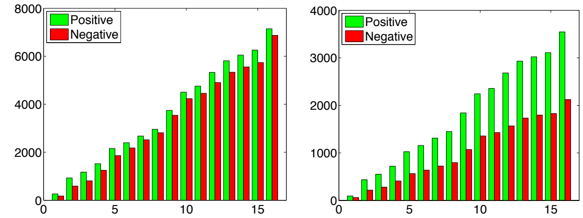
For example, for a two person group if the regret for person A is 0 and the regret for person B is 0.5 for viewing a program, it means that this group viewing event chose person A's favourite program to watch at this time and B had to compromise with

$$p(B's favourite program) - p(viewed program) = 0.5.$$

One drawback of this measure is that it is zero for a trivial model that always assigns probability zero to all items. This motivates another way to define regret, based on the log-loss in a ranking model:

$$l = -\log(p(i_v)) + \log\big(\sum_{i \in I} p(i)\big) \tag{8}$$

Regret definition (7) is more intuitive and it is constrained to lie between 0 and 1 which allows for better visualization of the regret



(a) Individual viewings (b) Group viewings

Figure 5: SpongeBob viewing statistics. Each pair the bar shows a program with the name SpongeBob. Since SpongeBob are broadcasted through different channels with different starting time.

measure. Log-loss regret is more principled and shows the predictive accuracy, however, the range is not bounded which makes the visualization not as easy to read as the first definition. Figure 7 shows the regret plots using both regret definitions for different type of size 2 groups. [4] The green dots are computed using only the individual MatchBox part of the Model to predict viewing probabilities whereas the blue dots represent probabilities predicted by the complete GroupBox model. We can see that GroupBox is able to fit the data better since the regret are more concentrated to the middle and origin of the plots. In addition, the regret plots show results that are consistent with the learned user influence. For example, in Figure 2 (b), a male adult's influence is much bigger than a male child's influence (around 2 times), and the corresponding regret plot in Figure 7 (b) shows that there more points on the male child's side of the plot which indicates that the male adult influences the group decision more. The same applies for all other cases.

7. DISCUSSION

In this paper we presented a probabilistic model called GroupBox, which is designed for making recommendations to groups of individuals. This model explicitly models users and items within both an individual and a group context, and also models influence of a user within a group. We applied this model to a large scale real world data set and showed how the results could provide us with a deeper understanding of group viewing activities. We validated that the various latent variables learned by the model had meaningful interpretations consistent with their design. Groupbox is a general model and is scalable to large data sets, and therefore we expect it to have wide applicabililty.

8. REFERENCES

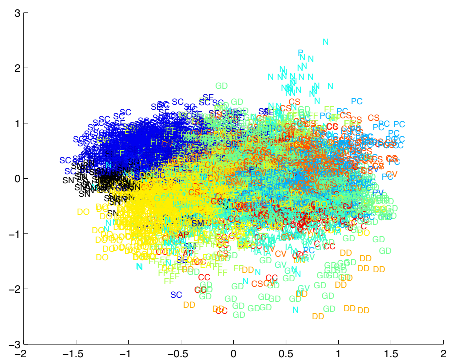
[1] S. Amer-Yahia, S. B. Roy, A. Chawlat, G. Das, and C. Yu. Group recommendation: Semantics and efficiency. *Proc. VLDB Endow.*, 2(1):754–765, August 2009.

[2] S. Seko andT. Yagi, M. Motegi, and S. Muto. Group recommendation using feature space representing behavioral tendency and power balance among members. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pages 101–108, 2011.

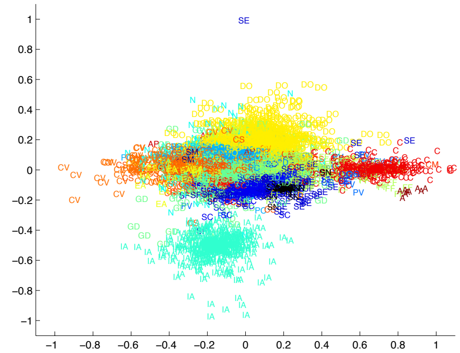
[3] E. Aronson, T. D. Wilson, R. M. Akert, and B. Fehr. *Social psychology*. Pearson Education Canada, 2012.

[4] L. Baltrunas, T. Makcinskas, and F. Ricci. Group recommendations with rank aggregation and collaborative filtering. In *Proceedings of the Fourth ACM Conference on*

[4] More regret plots are presented in the supplement.



(a) Item Traits



(b) Item Context Traits

Figure 6: The first 2 principle components of learned Item Traits and learned Item Context Traits. Programs are colour coded according to their type and marked with program type code from Table 2

Recommender Systems, RecSys '10, pages 119–126. ACM, 2010.

[5] S. Berkovsky, J. Freyne, and M. Coombe. Aggregation trade offs in family based recommendations. In *Proceedings of the 22Nd Australasian Joint Conference on Advances in Artificial Intelligence*, AI '09, pages 646–655. Springer-Verlag, 2009.

[6] L. Boratto, S. Carta, A. Chessa, M. Agelli, and M. L. Clemente. Group recommendation with automatic identification of users communities. In *Proceedings of the 2009 IEEE/WIC/ACM International Conference on Web Intelligence and International Conference on Intelligent Agent Technology - Workshops, Milan, Italy, 15-18 September 2009*, pages 547–550, 2009.

[7] A. J. B. Chaney, M. Gartrell, J. M. Hofman, J. Guiver, N. Koenigstein, P. Kohli, and U. Paquet. A large-scale exploration of group viewing patterns. In *TVX*, 2014.

[8] M. Gartrell, X. Xing, Q. Lv, A. Beach, R. Han, S. Mishra, and K. Seada. Enhancing group recommendation by incorporating social relationship interactions. In *Proceedings of the 16th ACM International Conference on Supporting Group Work*, GROUP '10, pages 97–106. ACM, 2010.

[9] A. Jameson and B. Smyth. The adaptive web. chapter Recommendation to Groups, pages 596–627. 2007.

[10] T. Minka, J.M. Winn, J.P. Guiver, S. Webster, Y. Zaykov, B. Yangel, A. Spengler, and J. Bronskill. Infer.NET 2.6. `http://research.microsoft.com/infernet`, 2014. Microsoft Research Cambridge.

[11] T. P. Minka. Expectation propagation for approximate bayesian inference. In *UAI*, 2001.

[12] U. Paquet and N. Koenigstein. One-class collaborative filtering with random graphs. In *WWW*, 2013.

[13] T. De Pessemier, S. Dooms, and L. Martens. Comparison of group recommendation algorithms. *Multimedia Tools and Applications*, 72:2497–2541, 2014.

[14] J. Recio-Garcia, G. Jimenez-Diaz, A. A. Sanchez-Ruiz, and B. Diaz-Agudo. Personality aware recommendations to groups. In *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09, pages 325–328. ACM, 2009.

[15] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 452–461, 2009.

[16] A. Said, S. Berkovsky, and E. W. De Luca. Group recommendation in context. In *Proceedings of the 2nd Challenge on Context-Aware Movie Recommendation*, pages 2–4. ACM, 2011.

[17] D. Stern, R. Herbrich, and T. Graepel. Matchbox: large scale online bayesian recommendations. In *WWW*, 2009.

[18] J. Winn and C. M. Bishop. Variational message passing. *Journal of Machine Learning Research*, 6:661–694, 2005.

[19] M. Ye, X. Liu, and W. C. Lee. Exploring social influence for recommendation: A generative model approach. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 671–680. ACM, 2012.

[20] Z. Yu, Z. Yu, X. Zhou, and Y. Nakamura. Handling conditional preferences in recommender systems. In *Proceedings of the 14th International Conference on Intelligent User Interfaces*, IUI '09, pages 407–412. ACM, 2009.

[21] Z. Yu, X. Zhou, Y. Hao, and J. Gu. Tv program recommendation for multiple viewers based on user profile merging. *User Modeling and User-Adapted Interaction*, 16(1):63–82, March 2006.

[22] Q. Yuan, G. Cong, and C. Y. Lin. COM: a generative model for group recommendation. In *KDD*, 2014.

Supplement

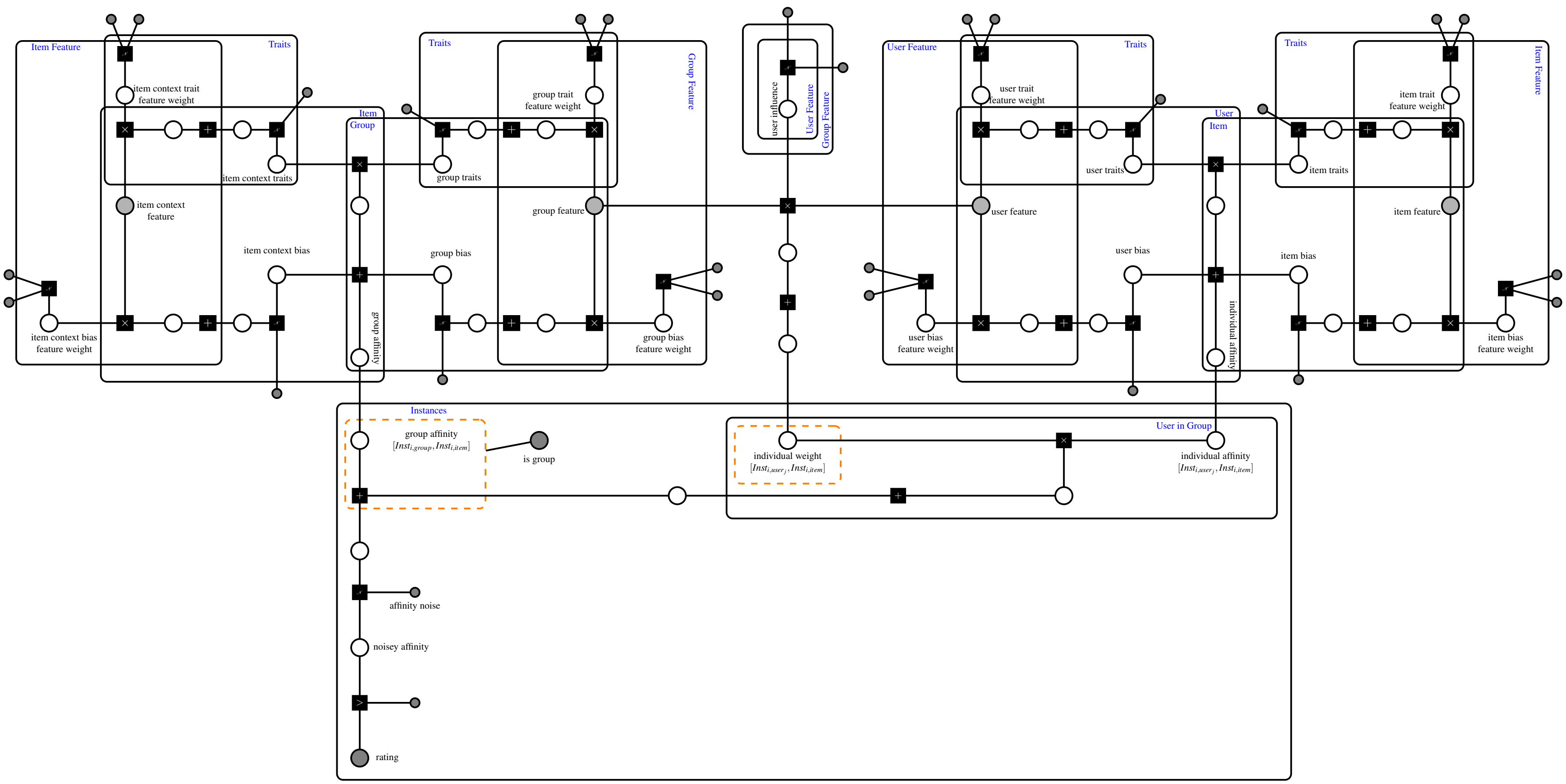
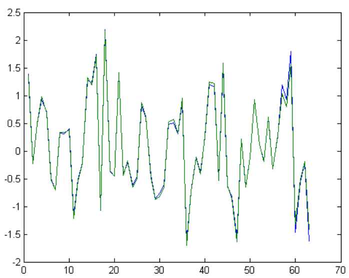


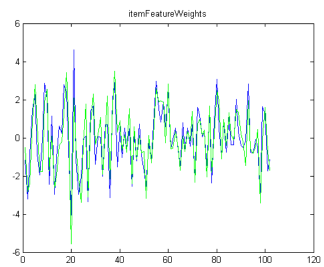
Figure 1: Factor Graph Representation of GroupBox

1	size 2 only female only child group	31	size 3 only female only child	64	size 4 only female only child
2	size 2 only female only youth group	32	size 3 only female only youth group	65	size 4 only female only youth group
3	size 2 only female only adult group	33	size 3 only female only adult group	66	size 4 only female only adult group
4	size 2 only female only senior group	34	size 3 only female only senior group	67	size 4 only female only senior group
5	size 2 only female child and youth group	35	size 3 only female child and youth group	68	size 4 only female child and youth group
6	size 2 only female child and adult mix group	36	size 3 only female child and adult mix group	69	size 4 only female child and adult mix group
7	size 2 only female child and senior mix group	37	size 3 only female child and senior mix group	70	size 4 only female child and senior mix group
8	size 2 only female youth and adult mix group	38	size 3 only female youth and adult mix group	71	size 4 only female youth and adult mix group
9	size 2 only female youth and senior mix group	39	size 3 only female youth and senior mix group	72	size 4 only female youth and senior mix group
10	size 2 only female adult and senior mix group	40	size 3 only female adult and senior mix group	73	size 4 only female adult and senior mix group
		41	size 3 only female other mix age mix group	74	size 4 only female other mix age mix group
11	size 2 only male only child group	42	size 3 only male only child	75	size 4 only male only child
12	size 2 only male only youth group	43	size 3 only male only youth group	76	size 4 only male only youth group
13	size 2 only male only adult group	44	size 3 only male only adult group	77	size 4 only male only adult group
14	size 2 only male only senior group	45	size 3 only male only senior group	78	size 4 only male only senior group
15	size 2 only male child and youth group	46	size 3 only male child and youth group	79	size 4 only male child and youth group
16	size 2 only male child and adult mix group	47	size 3 only male child and adult mix group	80	size 4 only male child and adult mix group
17	size 2 only male child and senior mix group	48	size 3 only male child and senior mix group	81	size 4 only male child and senior mix group
18	size 2 only male youth and adult mix group	49	size 3 only male youth and adult mix group	82	size 4 only male youth and adult mix group
19	size 2 only male youth and senior mix group	50	size 3 only male youth and senior mix group	83	size 4 only male youth and senior mix group
20	size 2 only male adult and senior mix group	51	size 3 only male adult and senior mix group	84	size 4 only male adult and senior mix group
		52	size 3 only male other mix age mix group	85	size 4 only male other mix age mix group
21	size 2 mix gender only child group	53	size 3 mix gender only child	86	size 4 mix gender only child
22	size 2 mix gender only youth group	54	size 3 mix gender only youth group	87	size 4 mix gender only youth group
23	size 2 mix gender only adult group	55	size 3 mix gender only adult group	88	size 4 mix gender only adult group
24	size 2 mix gender only senior group	56	size 3 mix gender only senior group	89	size 4 mix gender only senior group
25	size 2 mix gender child and youth group	57	size 3 mix gender child and youth group	90	size 4 mix gender child and youth group
26	size 2 mix gender child and adult mix group	58	size 3 mix gender child and adult mix group	91	size 4 mix gender child and adult mix group
27	size 2 mix gender child and senior mix group	59	size 3 mix gender child and senior mix group	92	size 4 mix gender child and senior mix group
28	size 2 mix gender youth and adult mix group	60	size 3 mix gender youth and adult mix group	93	size 4 mix gender youth and adult mix group
29	size 2 mix gender youth and senior mix group	61	size 3 mix gender youth and senior mix group	94	size 4 mix gender youth and senior mix group
30	size 2 mix gender adult and senior mix group	62	size 3 mix gender adult and senior mix group	95	size 4 mix gender adult and senior mix group
		63	size 3 mix gender other mix age mix group	96	size 4 mix gender other mix age mix group

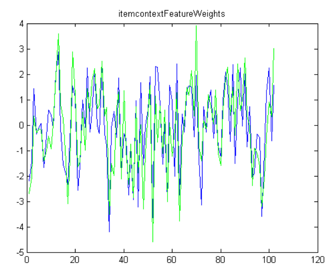
Table 1: Group Feature List



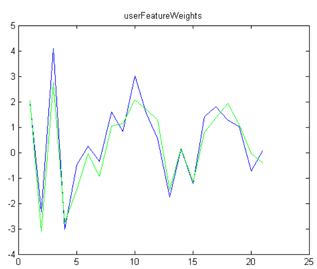
(a) User Influence



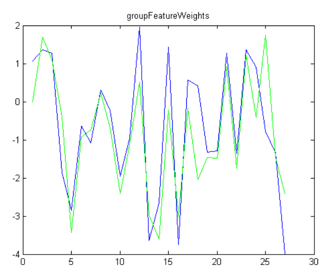
(b) Item Trait Feature Weights Estimation



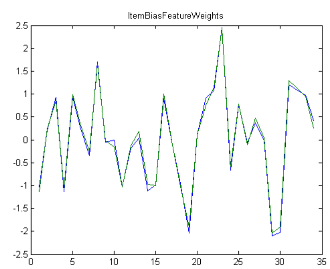
(c) Item Context Trait Feature Weights Estimation



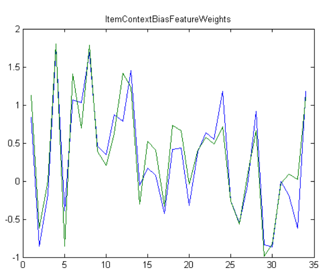
(d) User Feature Trait Weights Estimation



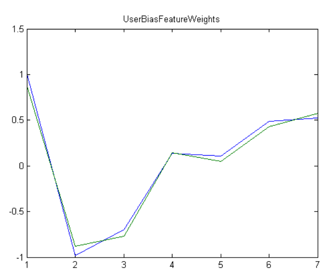
(e) Group Feature Trait Weights Estimation



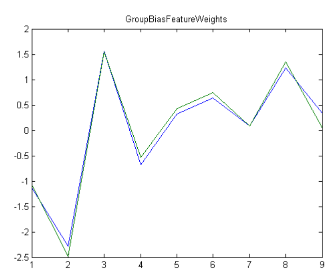
(f) Item Bias Feature Weights Estimation



(g) Item Context Bias Feature Weights Estimation



(h) User Bias Feature Weights Estimation



(i) Group Bias Feature Weights Estimation

	female. 0–11	female. 12–25	female. 26–38	female. 39–50	female. 51–62	female. 63–99	male. 0–11	male. 12–25	male. 26–38	male. 39–50	male. 51–62	male. 63–99
only female:Size2:Child(1–11)	1											
only female:Size2:Youth(12–25)		1										
only female:Size2:Adult(26–62)			0.29823	0.33927	0.36251							
only female:Size2:Senior(62–99)						1						
only female:Size2:Child & Youth	0.34074	0.65926										
only female:Size2:Child & Adult	0.15758		0.27852	0.29076	0.27314							
only female:Size2:Child & Senior	0.39417					0.60583						
only female:Size2:Youth & Adult		0.21469	0.22899	0.25837	0.29795							
only female:Size2:Youth & Senior		0.39083				0.60917						
only female:Size2:Adult & Senior			0.17653	0.2331	0.23634	0.35404						
only male:Size2:Child(1–11)							1					
only male:Size2:Youth(12–25)								1				
only male:Size2:Adult(26–62)									0.29292	0.32725	0.37983	
only male:Size2:Senior(62–99)												1
only male:Size2:Child & Youth							0.38957	0.61043				
only male:Size2:Child & Adult							0.1604		0.28436	0.27448	0.28076	
only male:Size2:Child & Senior							0.41943					0.58057
only male:Size2:Youth & Adult								0.20584	0.21613	0.26845	0.30958	
only male:Size2:Youth & Senior								0.38157				0.61843
only male:Size2:Adult & Senior									0.18267	0.32286	0.26367	0.2308
mix gender:Size2:Child(1–11)	0.41286						0.58714					
mix gender:Size2:Youth(12–25)		0.49478						0.50522				
mix gender:Size2:Adult(26–62)			0.13273	0.14544	0.16897				0.16615	0.18754	0.19918	
mix gender:Size2:Senior(62–99)						0.47282						0.52718
mix gender:Size2:Child & Youth	0.19968	0.28698					0.21548	0.29786				
mix gender:Size2:Child & Adult	0.059898		0.12894	0.13992	0.13628		0.080611		0.16509	0.1243	0.16495	
mix gender:Size2:Child & Senior	0.1501					0.29222	0.25342					0.30426
mix gender:Size2:Youth & Adult		0.090632	0.11779	0.13655	0.13982			0.1098	0.12519	0.13541	0.1448	
mix gender:Size2:Youth & Senior		0.15945				0.34753		0.18693				0.30609
mix gender:Size2:Adult & Senior			0.053038	0.10032	0.1188	0.14473			0.13636	0.14109	0.14032	0.16534

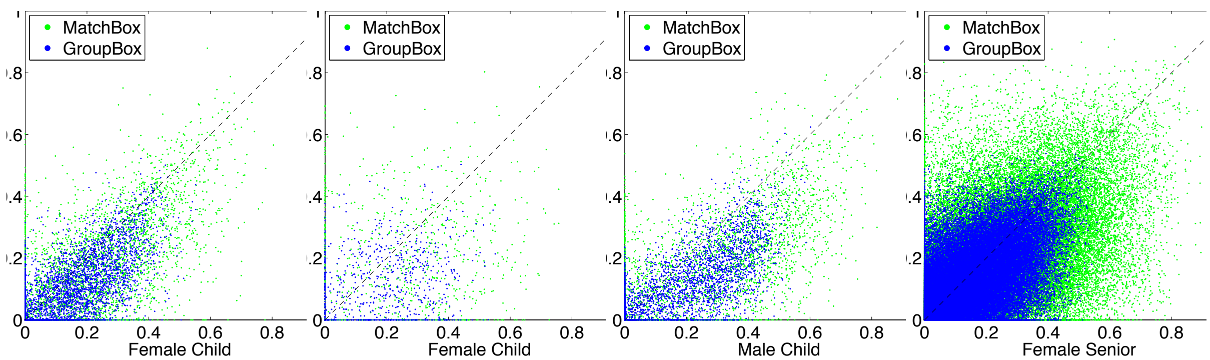
Figure 2: Size 2 Group Learned User Influence

	female. 0–11	female. 12–25	female. 26–38	female. 39–50	female. 51–62	female. 63–99	male. 0–11	male. 12–25	male. 26–38	male. 39–50	male. 51–62	male. 63–99
only female:Size3:Child(1–11)	1											
only female:Size3:Youth(12–25)		1										
only female:Size3:Adult(26–62)			0.27238	0.32311	0.40452							
only female:Size3:Senior(62–99)						1						
only female:Size3:Child & Youth	0.41549	0.58451										
only female:Size3:Child & Adult	0.11413		0.34274	0.31266	0.23046							
only female:Size3:Child & Senior	0.26808					0.73192						
only female:Size3:Youth & Adult		0.17653	0.26674	0.26711	0.28962							
only female:Size3:Youth & Senior		0.2458				0.7542						
only female:Size3:Adult & Senior			0.18575	0.22913	0.25742	0.32771						
only female:Size3:Mix	0.077507	0.16034	0.1283	0.21692	0.19513	0.22179						
only male:Size3:Child(1–11)							1					
only male:Size3:Youth(12–25)								1				
only male:Size3:Adult(26–62)									0.31274	0.34561	0.34165	
only male:Size3:Senior(62–99)												1
only male:Size3:Child & Youth							0.41202	0.58798				
only male:Size3:Child & Adult							0.11689		0.35153	0.31017	0.2214	
only male:Size3:Child & Senior							0.44343					0.55657
only male:Size3:Youth & Adult								0.15703	0.23982	0.3369	0.26626	
only male:Size3:Youth & Senior								0.41358				0.58642
only male:Size3:Adult & Senior									0.25548	0.11135	0.297	0.33616
only male:Size3:Mix							0.087881	0.15762	0.15458	0.20506	0.14231	0.25255
mix gender:Size3:Child(1–11)	0.45584						0.54416					
mix gender:Size3:Youth(12–25)		0.50618						0.49382				
mix gender:Size3:Adult(26–62)			0.12628	0.098209	0.14943				0.19582	0.20917	0.2211	
mix gender:Size3:Senior(62–99)						0.47023						0.52977
mix gender:Size3:Child & Youth	0.21422	0.31238					0.17763	0.29577				
mix gender:Size3:Child & Adult	0.062124		0.11785	0.1235	0.14043		0.069869		0.1542	0.16669	0.16534	
mix gender:Size3:Child & Senior	0.13458					0.25462	0.243					0.3678
mix gender:Size3:Youth & Adult		0.074195	0.10204	0.11992	0.12342			0.10114	0.15051	0.1537	0.17507	
mix gender:Size3:Youth & Senior		0.14001				0.34953		0.26397				0.24648
mix gender:Size3:Adult & Senior			0.070343	0.084203	0.12643	0.15706			0.1276	0.1348	0.11763	0.18194
mix gender:Size3:Mix	0.048517	0.062066	0.077487	0.098228	0.10131	0.093692	0.047612	0.067986	0.098251	0.10669	0.10456	0.093596

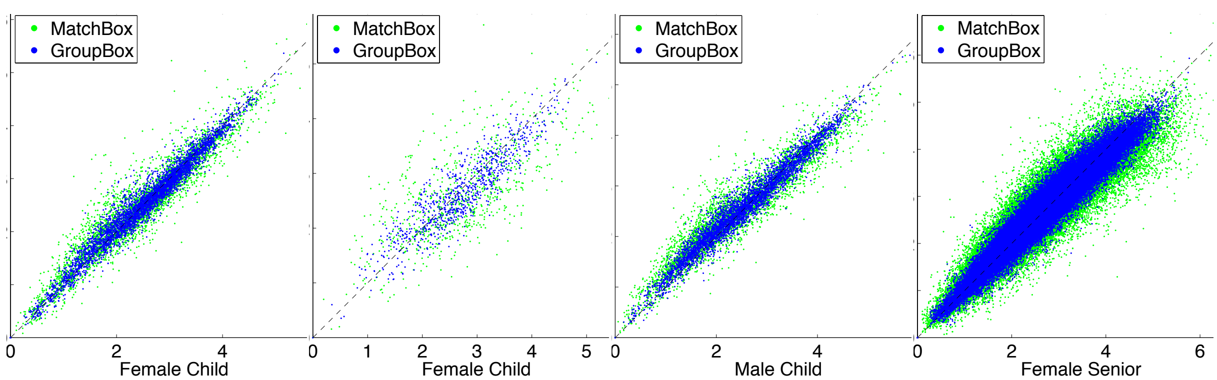
Figure 3: Size 3 Group Learned User Influence

	female. 0–11	female. 12–25	female. 26–38	female. 39–50	female. 51–62	female. 63–99	male. 0–11	male. 12–25	male. 26–38	male. 39–50	male. 51–62	male. 63–99
only female:Size4:Child(1–11)	1											
only female:Size4:Youth(12–25)		1										
only female:Size4:Adult(26–62)			0.84011	0.49033	−0.33043							
only female:Size4:Senior(62–99)						1						
only female:Size4:Child & Youth	0.3738	0.6262										
only female:Size4:Child & Adult	0.16099		0.28111	0.20015	0.35775							
only female:Size4:Child & Senior	0.29868					0.70132						
only female:Size4:Youth & Adult		0.12255	0.16934	0.27516	0.43296							
only female:Size4:Youth & Senior												
only female:Size4:Adult & Senior			0.16272	0.020932	0.33083	0.48552						
only female:Size4:Full Mix	0.11297	0.15511	0.17571	0.27055	0.14695	0.13872						
only male:Size4:Child(1–11)							1					
only male:Size4:Youth(12–25)								1				
only male:Size4:Adult(26–62)									0.35775	0.31399	0.32826	
only male:Size4:Senior(62–99)												
only male:Size4:Child & Youth							0.39229	0.60771				
only male:Size4:Child & Adult							0.13249		0.29683	0.28494	0.28573	
only male:Size4:Child & Senior							0.88535					0.11465
only male:Size4:Youth & Adult								0.18962	0.25477	0.239	0.31661	
only male:Size4:Youth & Senior								0.4847				0.5153
only male:Size4:Adult & Senior									0.18544		0.44629	0.36827
only male:Size4:Full Mix							0.10896	0.11918	0.20825	0.17019	0.13465	0.25878
mix gender:Size4:Child(1–11)	0.47642						0.52358					
mix gender:Size4:Youth(12–25)		0.46963						0.53037				
mix gender:Size4:Adult(26–62)			0.07254	0.020623	0.075022				0.23582	0.17885	0.41715	
mix gender:Size4:Senior(62–99)						0.51135						0.48865
mix gender:Size4:Child & Youth	0.1303	0.37755					0.15399	0.33816				
mix gender:Size4:Child & Adult	0.037927		0.12313	0.11529	0.13797		0.06564		0.18151	0.14395	0.19457	
mix gender:Size4:Child & Senior	0.17302					0.22488	0.22422					0.37788
mix gender:Size4:Youth & Adult		0.06488	0.12786	0.099039	0.11508			0.088676	0.14712	0.15881	0.19853	
mix gender:Size4:Youth & Senior		0.26121				0.21716		0.26132				0.2603
mix gender:Size4:Adult & Senior			0.03787	0.0015279	0.088559	0.19236			0.24564	0.13995	0.1695	0.1246
mix gender:Size4:Full Mix	0.046056	0.05565	0.072753	0.099086	0.091015	0.083261	0.043755	0.070124	0.11531	0.11127	0.1099	0.10182

Figure 4: Size 4 Group Learned User Influence



(a) Regret plot; Group type 5 (b) Regret plot; Group type 7 (c) Regret plot; Group type 15 (d) Regret plot; Group type 24



(e) Log-Loss Regret; Group type 5 (f) Log-Loss Regret; Group type 7 (g) Log-Loss Regret; Group type 15 (h) Log-Loss Regret; Group type 24

Figure 5: More examples of regret plot for group types with 2 types of definition.

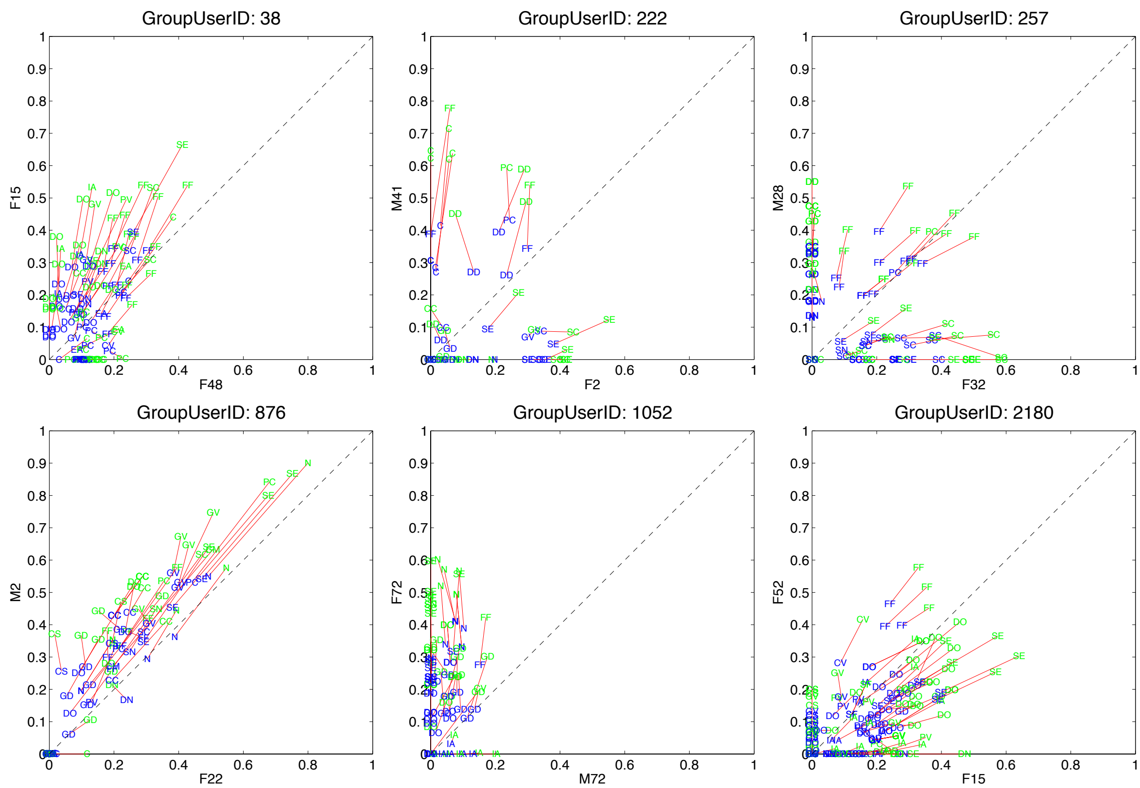


Figure 6: Examples of regret plots for a group

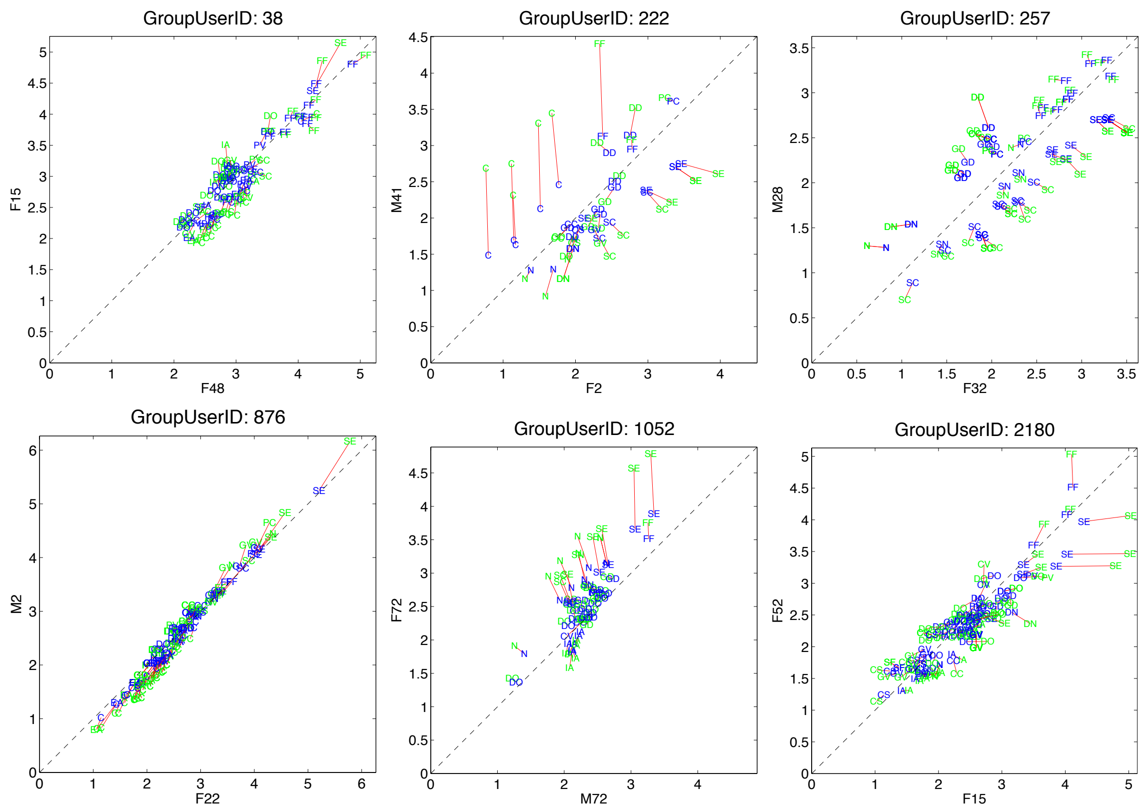


Figure 7: Examples of ranking log loss based regret plots for a group