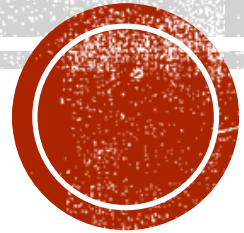
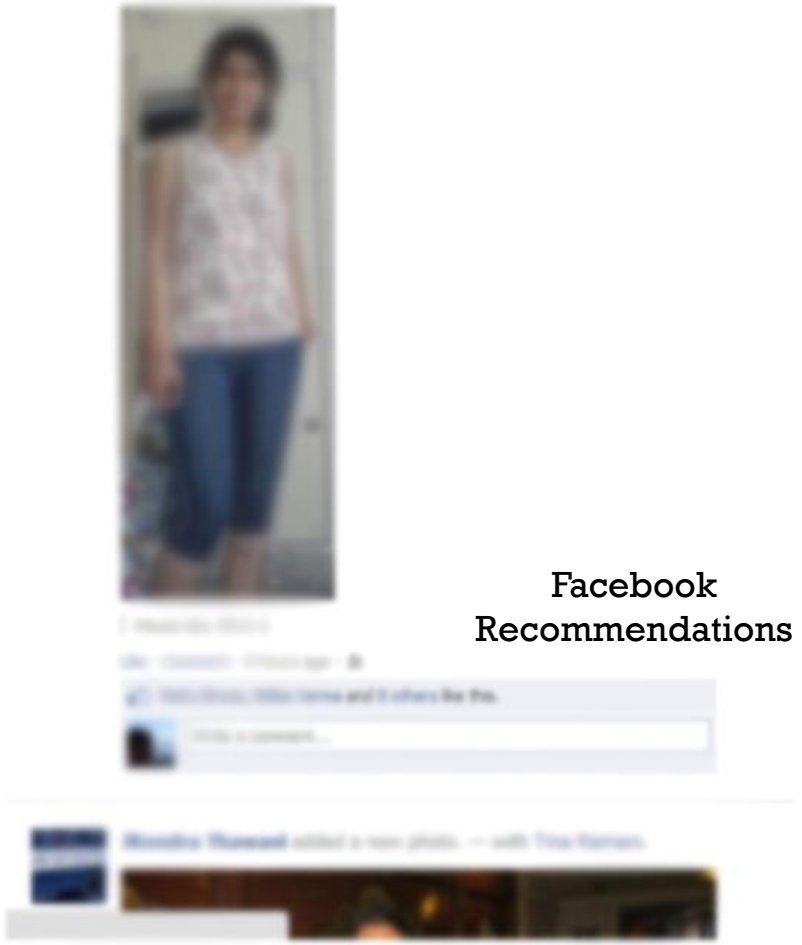
RECOMMENDATION SYSTEM
IN SOCIAL NETWORKS



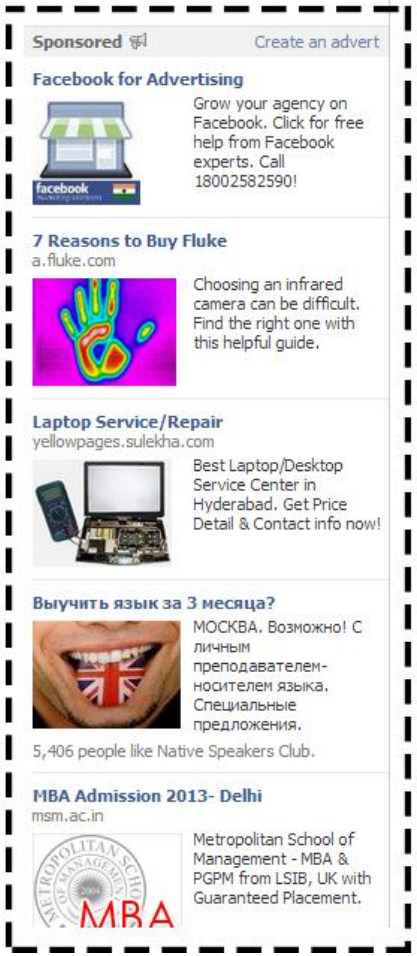
Presented by
Abhishek Gupta (2010EE10435)
Parijat Mazumdar (2010EE10467)

MOTIVATION

Social Networks



Facebook Recommendations



PROBLEM DEFINITION

Item **j**

User **i**

(i,j): response y_{ij}

Which item should we select?

- The one with highest predicted CTR ------ **EXPLOIT**
- The one most useful for the CTR prediction model -------**EXPLORE**

DATASET

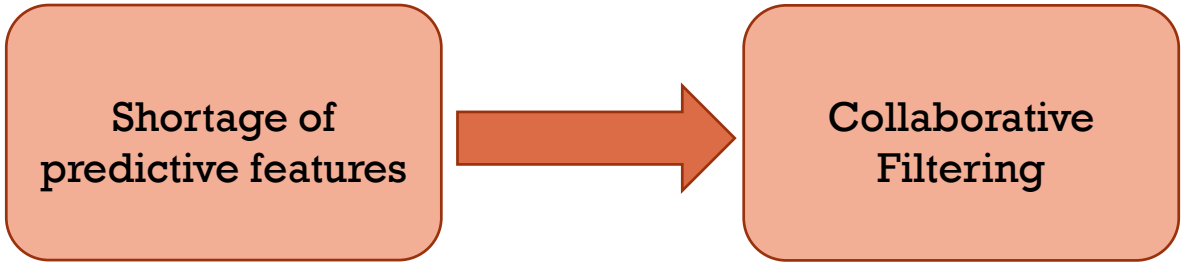
- Raw Training Data :
  - 70M records spanning 31 days
- User profile data:
  - Tags
  - Keywords
  - Date of Birth (Age)
  - Gender
- User action data:
  - Tweet
  - Retweet
  - Comment

UserID	ItemID	Choice	Timestamp
100001	372259	0	45237
137136	458026	1	49922
⋮	⋮	⋮	⋮

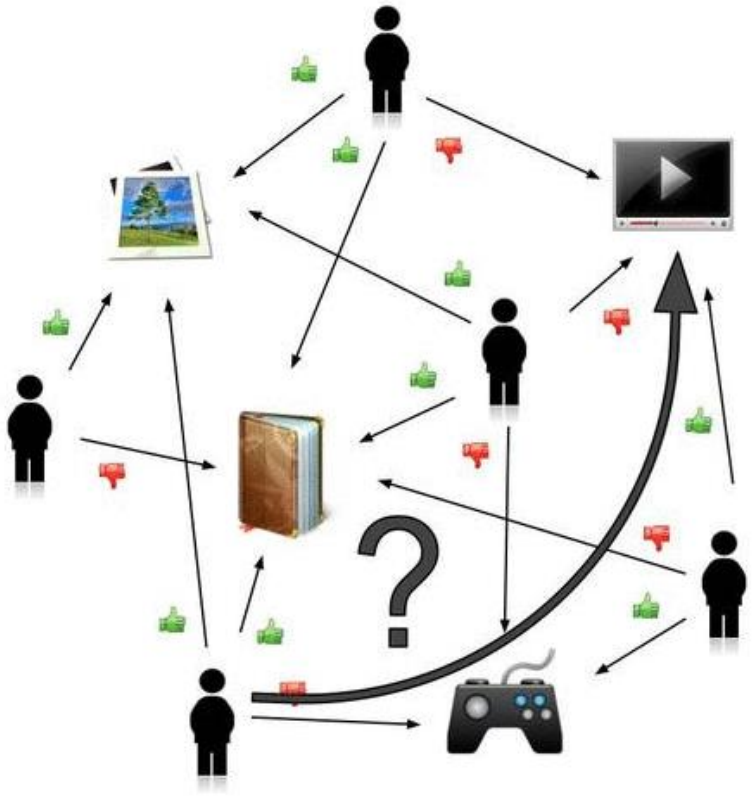
	Id_1	.	458026	.	Id_n
UId_1	.	.	.	.	.
⋮	⋮	⋮	⋮	⋮	⋮
137136	.	.	1	.	.
⋮	⋮	⋮	⋮	⋮	⋮
UId_n	.	.	.	.	.

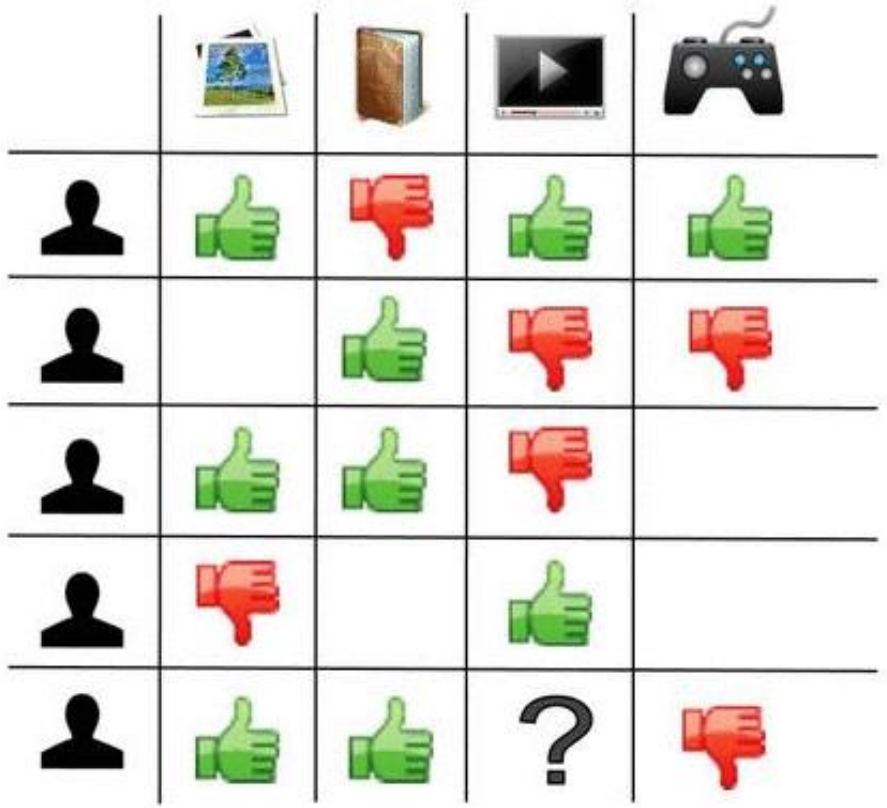
SALIENT FEATURES OF DATASET

- 1.3 million users, 4,700 items
- **99.31%** sparse
- Many users haven't rated a single item
- Many items not rated
- **45%** users' tags missing **47%** users' keywords missing
- Gender not known of roughly **20%** users



COLLABORATIVE FILTERING





PRE-PROCESSING

- Duplicates Removal
  - 41M records remain
- Removing Dummy users and items
  - Removed 10% of the highly unrated movies to reduce sparsity
  - Removed 0.00001% of the users with very few ratings

	1	2	3	4	5	.....	4278	4279
100001	1	0	1	1	0	.....	1	0
100002	0	0	1	0	0	.....	0	0
. . . . .	. . . . .		. . . . .	. . . . .		....	. . . . .	. . . . .
103792	1	0	1	1	0	.....	1	1
103793	0	0	0	0	0	.....	1	0
. . . . .	. . . . .		. . . . .	. . . . .		....	. . . . .	. . . . .

RATING-BIAS BASED MODEL

- Item bias: e.g. Titanic being a classic movie has positive bias
- User bias: Some users critical, others lenient

$$\widehat{r_{ij}} = bu_i + bi_j + \mu$$

Predicted rating

Global Bias

$$bi_j = \sum_{j \in R(i)} \frac{(r_{ij} - \mu)}{|R(i)|}$$

$$bu_i = \sum_{i \in R(j)} \frac{(r_{ij} - b_i - \mu)}{|R(j)|}$$

	1	3	4	.....	j	.....	4278	4279
100001	1	1	1	.....	0	.....	1	0
⋮	⋮	⋮	⋮	.....	⋮	.....	⋮	⋮
i	0	1	1	.....	1	.....	0	1
⋮	⋮	⋮	⋮	.....	⋮	.....	⋮	⋮
103792	1	1	1	.....	0	.....	1	1
⋮	⋮	⋮	⋮	.....	⋮	.....	⋮	⋮

$$bu_i = avg\{A_{ij} - bi_j - bg\}$$

$$bi_j = avg\{A_{.j}\} - bg$$

- $bg = avg\{A_{ij}\} = -0.7488$

Global Bias

REGULARIZATION

$$bi_j = \sum_{j \in R(i)} \frac{(r_{ij} - \mu)}{\lambda_j + |R(i)|}$$

$$bu_i = \sum_{i \in R(j)} \frac{(r_{ij} - b_i - \mu)}{\lambda_i + |R(j)|}$$

Introducing regularization terms

Taking $\lambda_u = 0$

λ_i	MAP@3
0	0.2612
100	0.2705
500	0.2694

Taking $\lambda_i = 100$

λ_u	MAP@3
0	0.2705
100	0.2784
500	0.2833
1000	0.2815

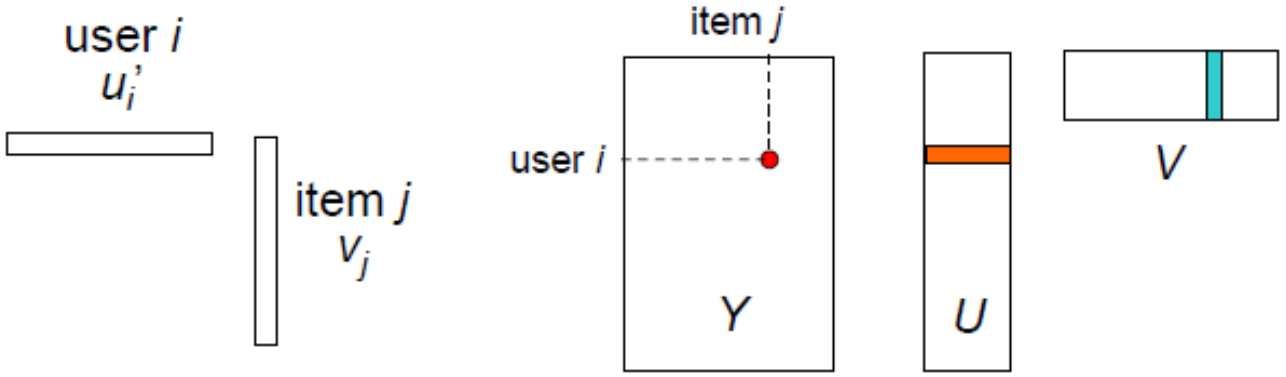
CAPTURING USER-ITEM INTERACTIONS

- Personalized recommendations

$$\widehat{r_{ij}} = bu_i + bi_j + \mu + p_i^T q_j$$

User latent factors Dim: 20

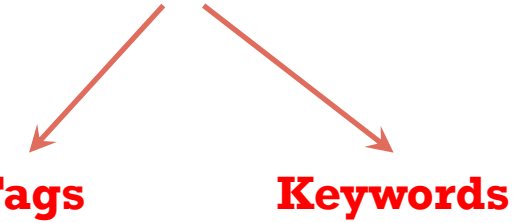
Item latent factors Dim: 20



Λ	MAP@3
100	0.2857
500	0.2865
1000	0.2871
5000	0.2868

SIGMA-TAG MODEL

- Each user is a combination of his interests!



$$\widehat{r_{ij}} = bu_i + bi_j + \mu + (\sum_{k \in T(i)} t_k + \sum_{l \in K(i)} w_{il} k_l)^T q$$

- Number of Tags = 130304 ⇒ Unknowns reduced immensely
- **What about users without tags/ keywords?**

USING SOCIAL NETWORK DATA

Id_i	Id_j	3	6	5

Number of times user **i** has "@" user **j**

Number of times user **i** retweeted user **j**

Number of times user **i** commented user **j**

	Id_1	.	Id_j	.	Id_n
Id_1	.	.	.	.	.
.	.	.	.	.	.
Id_i	.	.	14	.	.
.	.	.	.	.	.
Id_n	.	.	.	.	.

$$p_i = \sum_{k \in ID(i)} M_{ik} * p_k$$

$$\widehat{r_{ij}} = \sum_{k \in ID(i)} M_{ik} * \widehat{r_{kj}}$$

INTRODUCING AGE & GENDER BIAS

- Gender Bias
  - $b_{gen(u),i}$ ⟶ Biasness of user's gender towards the particular item

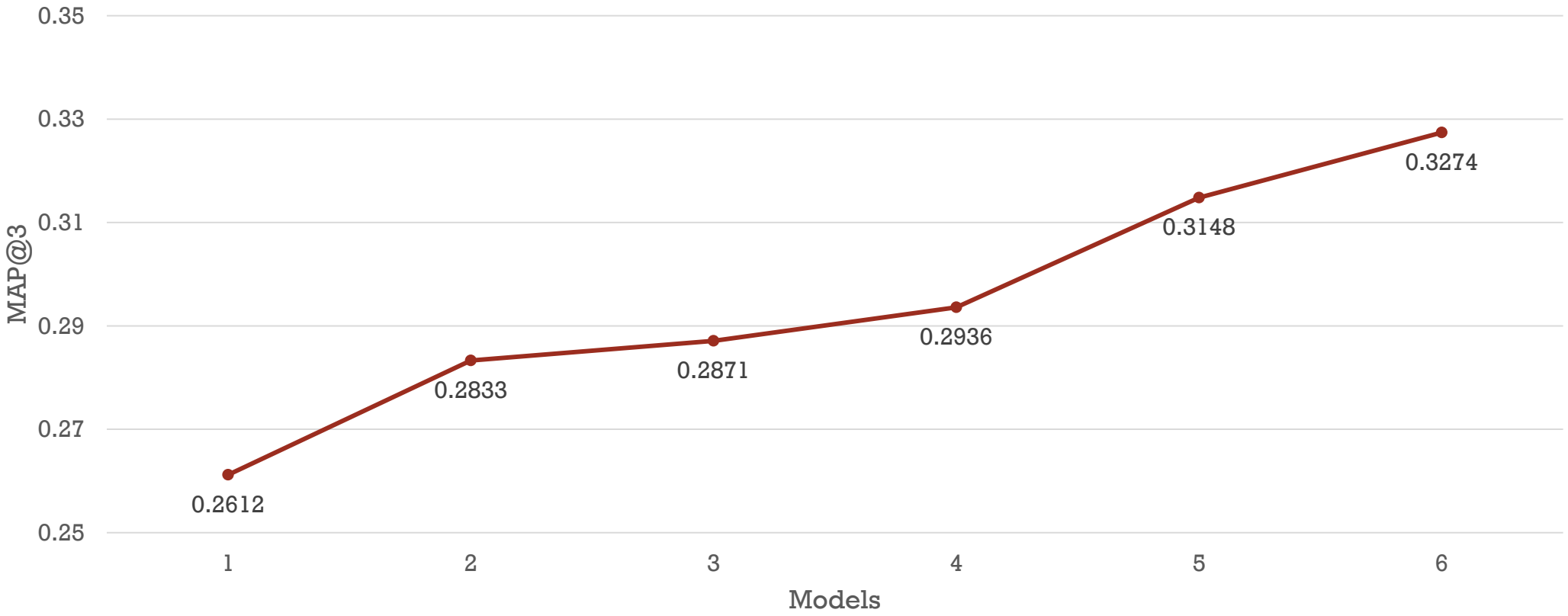
$$b_{gen(u),i} = \begin{cases} \sum_{k \in M} A_{ki}/_{n} & \text{if user's gender is male} \\ \sum_{k \in F} A_{ki}/_{n} & \text{if user's gender if female} \\ \sum_{k \in U} A_{ki}/_{n} & \text{if user's gender is unknown} \end{cases}$$

- Age Bias
  - $b_{age(u),i}$ ⟶ Biasness of user's age towards the particular item
  - For age bias we have created 30 bins and set the age bias accordingly

$$\widehat{r_{ij}} = \boldsymbol{bu_i} + \boldsymbol{bi_j} + \boldsymbol{b_{gen(i),j}} + \boldsymbol{b_{age(i),j}} + \boldsymbol{\mu} + (\sum_{\boldsymbol{k \in T(i)}} \boldsymbol{t_k} + \sum_{\boldsymbol{l \in K(i)}} \boldsymbol{w_{il} k_l})^{\boldsymbol{T}} \boldsymbol{q}$$

PROGRESS SUMMARY

Model	MAP@3
Bias only	0.2612
Bias with regularization	0.2833
Basic latent factor model	0.2871
Sigma Tag model	0.2936
Sigma tag + social network data	0.3148
Sigma tag + age, gender bias	0.3274


0.35
0.33
0.31
0.29
0.27
0.25
MAP@3
0.2612
0.2833
0.2871
0.2936
0.3148
0.3274
1
2
3
4
5
6
Models

THANK YOU!!!