E-commercial Product Recommendations

Team Members		Expected Contributions*
Danning Sui	ds3516	Modeling, Data Processing, Algorithm Implementation
Panpan Cheng	pc2756	Modeling, Algorithm Implementation, Visualization
Qiong Hu	qh2174	Modeling, Data Processing, Algorithm Implementation

*Equally distributed the tasks and collaboratively completed the project.

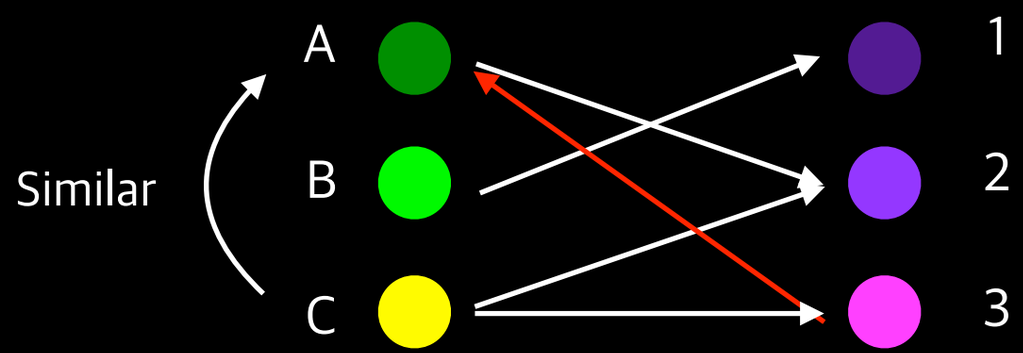
Overview

- Combination of User-based Collaborative Filtering and Item to Item Recommendation Process
- Pursuing a personalized and abundant recommendation set to open the market

User-based CF

Input: A Bipartite of user and product

According to their rating reviews, find the favored products from similar users.

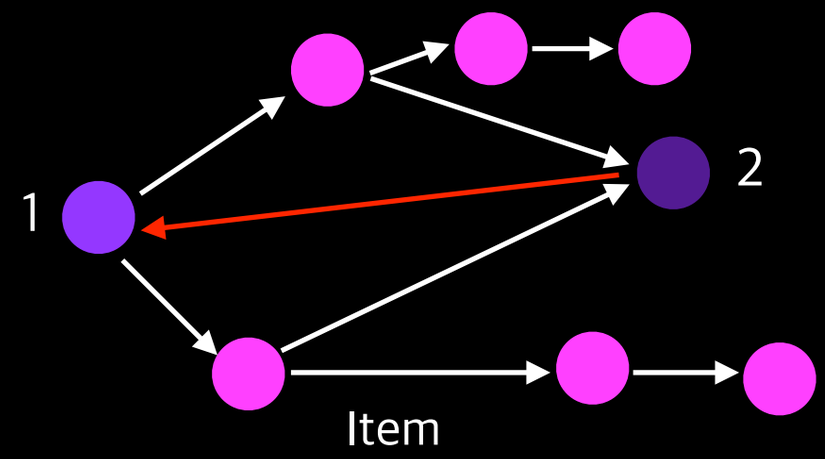


Outcome: Recommend 3 to A

Item to Item Recommendation

Input: A network of products

According to their co-purchasing network, compute their similarities and recommend.



Outcome: Recommend 2 to 1

Model and Algorithms

Input | Algorithms | Output

Part 1

User Review Rating Dataset →

User-based Collaborative Filtering (pySpark)

A. Load and parse the data
B. Build the recommendation model (Alternating Least Squares)
C. Evaluate the model on training data
D. Implement the model on testing data

→ Recommended Items for Each User (I)

Part 2

Co-purchasing Product Network →

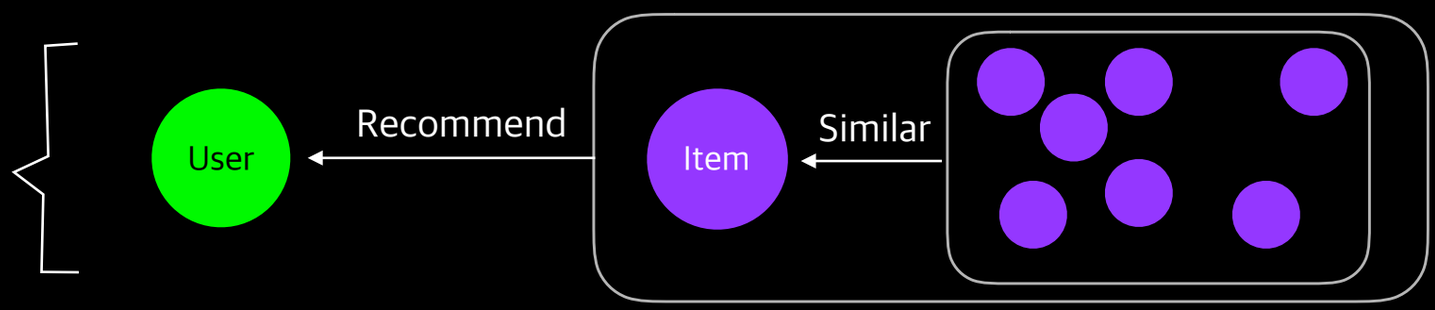
Item to Item Similarity (System G)

A. Load data and Visualize
B. Analyze the Graph by gShell Query
C. Find the similar sets by Colfilter function

→ Similar Items for Each Item (II)

Part 3

Output (I) and (II)



Datasets (1.14GB)

A. User Review Rating Matrix - 977.5MB

(amazon-meta.txt)

Dataset statistics	
Products	548,552
Product-Project Edges	1,788,725
Reviews	7,781,990
Product category memberships	2,509,699

Products by product group	
Books	393561
DVDs	19828
Music CDs	103144
Videos	26132

```
Id:   1
ASIN: 0827229534
  title: Patterns of Preaching: A Sermon Sampler
  group: Book
  salesrank: 396585
  similar: 5  0804215715  156101074X  0687023955  0687074231  082721619X
  categories: 2
   |Books[283155]|Subjects[1000]|Religion & Spirituality[22]|Christianity[12290]|Clergy[12360]|
Preaching[12368]
   |Books[283155]|Subjects[1000]|Religion & Spirituality[22]|Christianity[12290]|Clergy[12360]|
Sermons[12370]
  reviews: total: 2  downloaded: 2  avg rating: 5
    2000-7-28  cutomer: A2JW67OY8U6HHK  rating: 5  votes:  10  helpful:   9
    2003-12-14  cutomer: A2VE83MZF98ITY  rating: 5  votes:   6  helpful:   5
```

Datasets (1.14GB)

B. Co-purchasing Item Network - 158.1MB in total

(Amazon0302.txt, Amazon0312.txt, Amazon0505.txt, Amazon0601.txt)

Example - Amazon0312.txt

Dataset statistics	
Nodes	262111
Edges	1234877
Nodes in largest WCC	262111 (1.000)
Edges in largest WCC	1234877 (1.000)
Nodes in largest SCC	241761 (0.922)
Edges in largest SCC	1131217 (0.916)
Average clustering coefficient	0.4198
Number of triangles	717719
Fraction of closed triangles	0.09339
Diameter (longest shortest path)	32
90-percentile effective diameter	11

```
# Directed graph (each unordered pair of nodes is saved once): Amazon0312.txt
# Amazon product co-purchaisng network from March 12 2003
# Nodes: 400727 Edges: 3200440
# FromNodeId	ToNodeId
0	1
0	2
0	3
0	4
0	5
1	0
1	2
1	13
1	14
1	15
2	0
2	1
2	4
2	5
2	16
3	70
5	6
```

Outcomes

Model Part 1 : User-based Collaborative Filtering (pySpark)

Rank	User Id	Item Id	Original Score	Model Score	Error
#1	1080592	284560	5	4.99797677899322	0.00202322100678
#2	354796	227866	5	4.99583012902760	0.00416987097240
#3		488815	5	4.99583012902760	0.00416987097240
#4		54687	5	5.00331265792859	0.00331265792859
#5	1056524	274052	5	5.00026234076471	0.00026234076471
#6		500789	5	4.99341921309295	0.00658078690705
#7		354097	5	4.99341921309295	0.00658078690705
#8	199516	434236	5	4.99745108641090	0.00254891358910
#9		329445	5	4.99745108641090	0.00254891358910
#10		453871	5	4.99745108641090	0.00254891358910

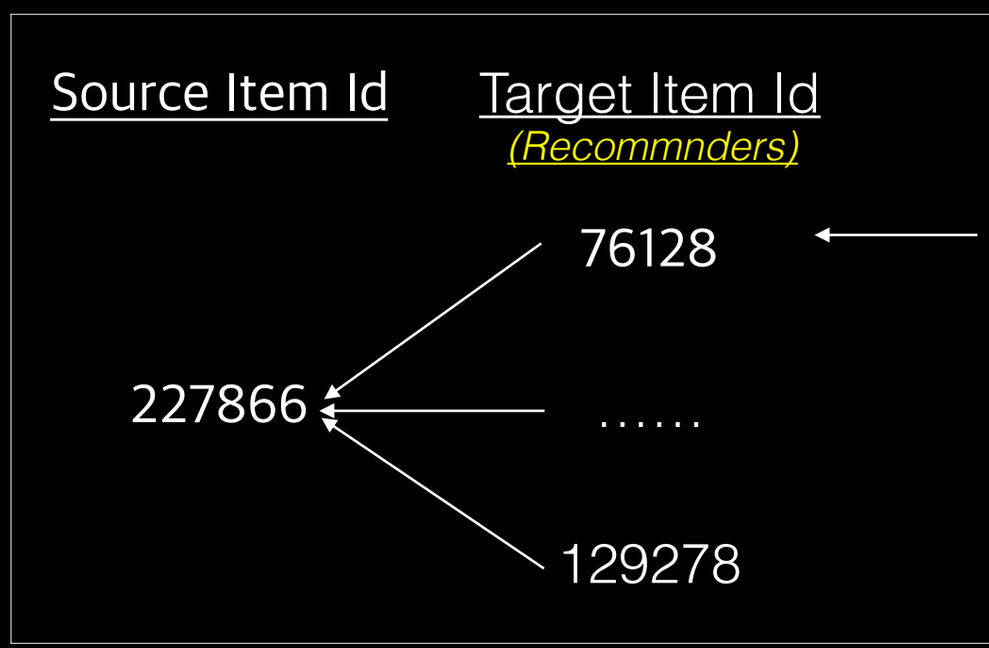
Example

Outcomes

Model Part 2 : Item to Item Similarity (System G) - depth=4

Rank	#Paths	Source Item Id	Target Item Id	Name	Category
#1	4	227866	76128	The Hunted	Book
#2	3		206319	Eating and Drinking in Paris: French Menu Reader and Restaurant Guide	Book
#3	3		166795	Rembrandt's Eyes	Book
#4	2		244665	The Complete Illustrated Guide to Aromatherapy	Book
#5	2		244664	Green Mansions	Video
#6	2		244663	Good Neighbors: Communicating With the Mexicans	Book
#7	2		244662	The Strategy Machine: Building Your Business One Idea at a Time	Book
#8	2		177920	The Century-America's Time	Video
#9	2		217435	The Cutout	Book
#10	2		129278	Adventure in Business: An I.M.M.E.R.S.I.O.N Approach to Training & Consulting	Book

Example

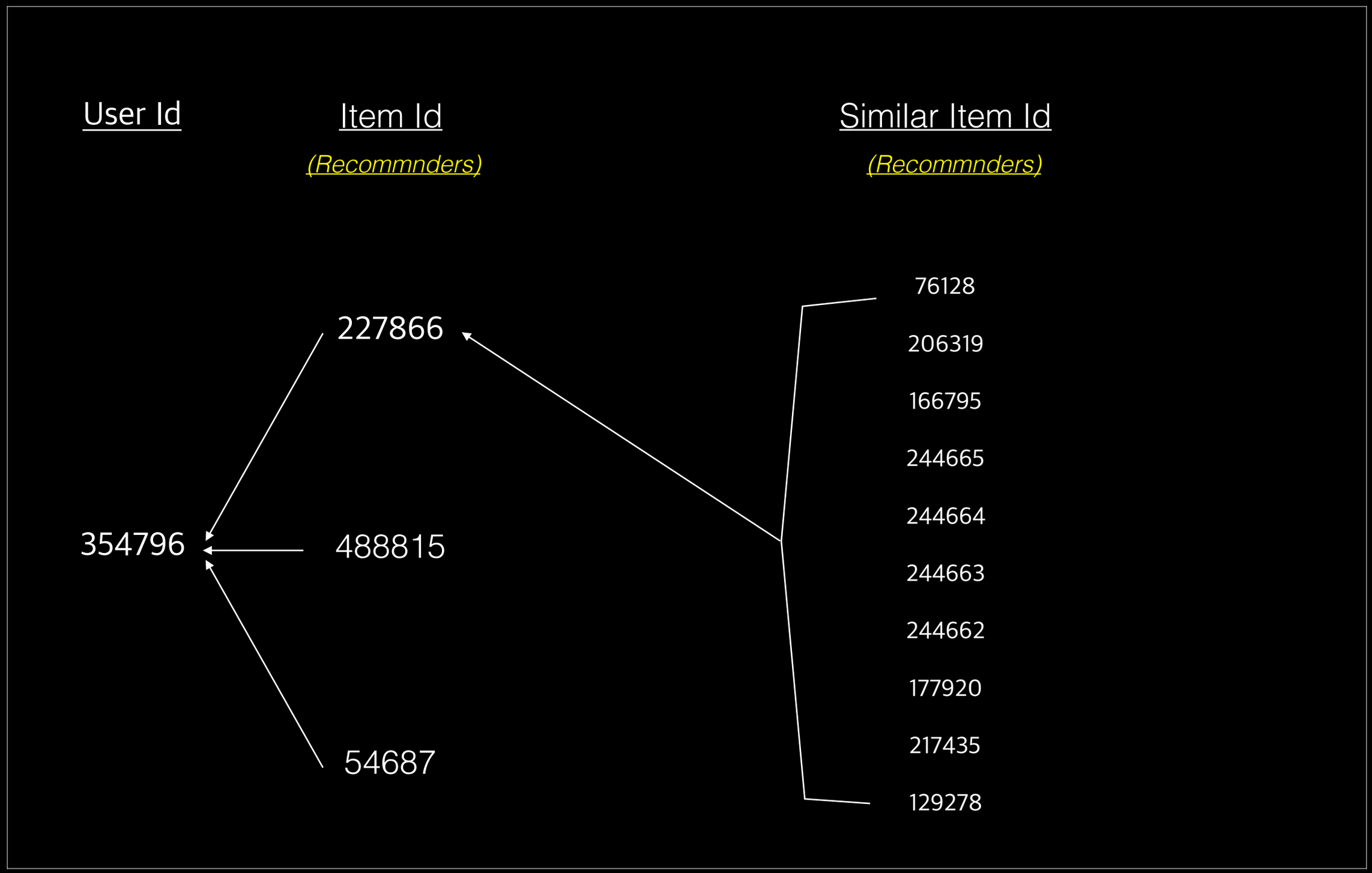


Detailed Info

ASIN: 0440134250
title: The Hunted
group: Book
salesrank: 770630
similar: 5 0060089547 0060082216 0380822253 0060082208 0060083999
categories: 3
|Literature & Fiction[17]|General[10125]|Contemporary[10129]
|Mystery & Thrillers[18]|Authors, A-Z[69300]|(L)[69312]|Leonard, Elmore[69519]|
General[284558]
|Mystery & Thrillers[18]|Authors, A-Z[69300]|(L)[69312]|Leonard, Elmore[69519]|
Paperback[285805]
reviews: total: 8 downloaded: 8 avg rating: 3.5

Outcomes

Model Part 3 : Recommendation Chain

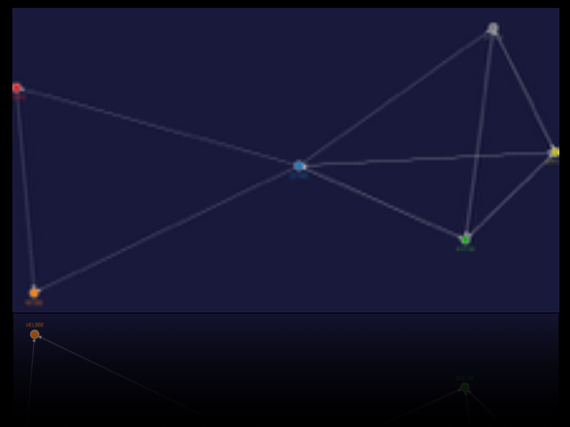


Outcomes

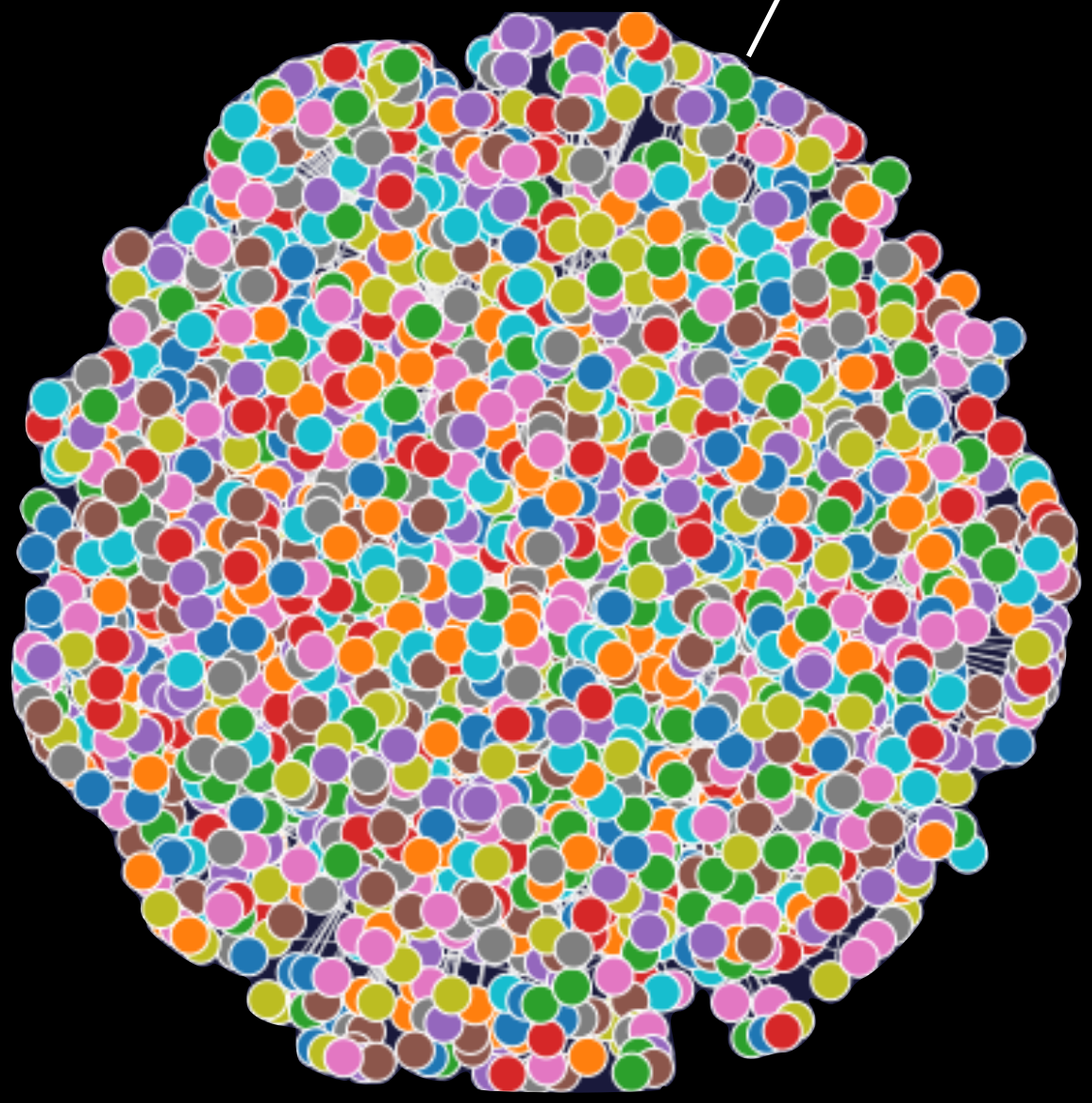
Model Part 2 Visualization

Neighbor Graph
(depth=1)

#Nodes: 5
#Edges: 9

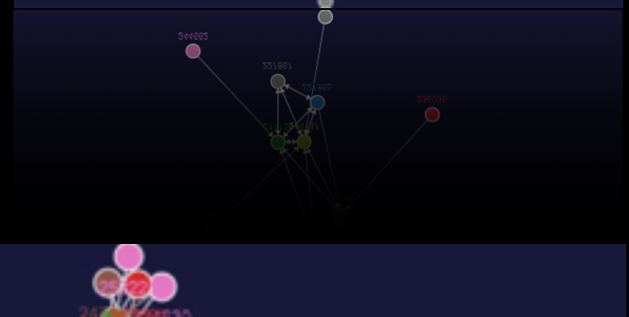


id=227866



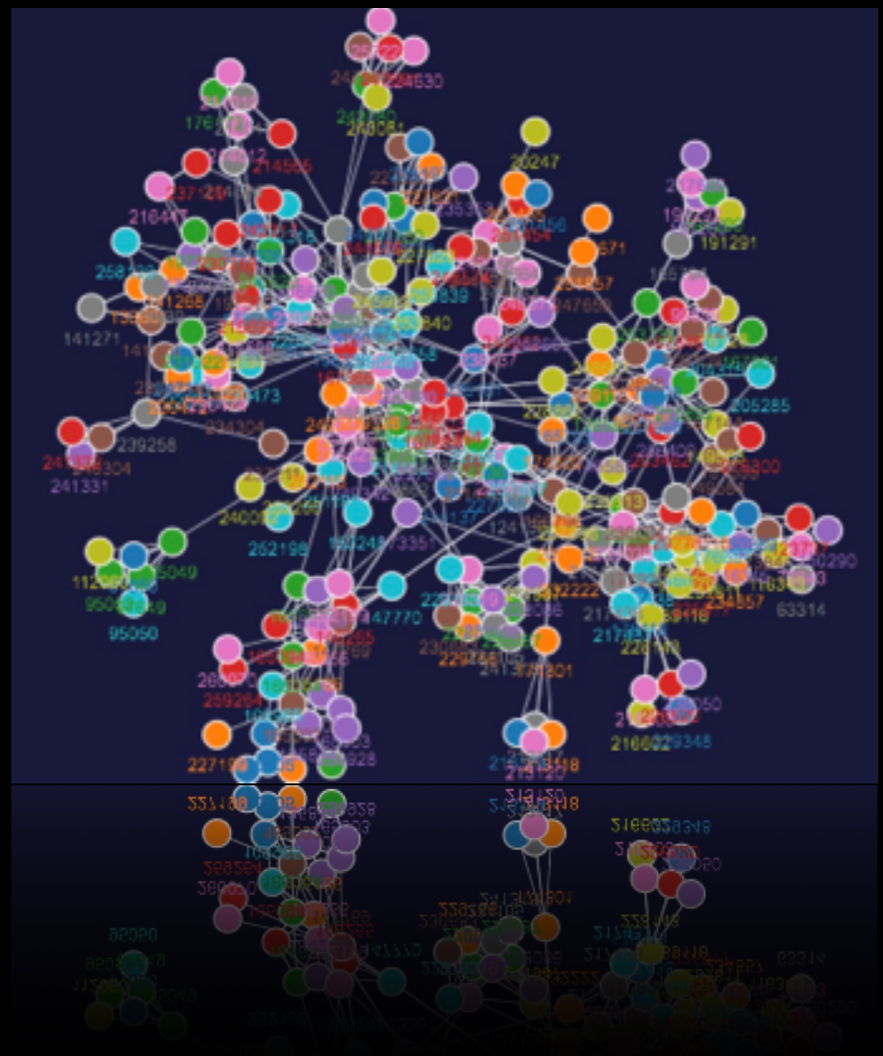
Neighbor Graph
(depth=2)

#Nodes:49
#Edges:169



Neighbor Graph
(depth=4)

#Nodes: 245
#Edges: 969



Challenges

- Large Data Size Process Efficiency

- Fitness of CF model
  - ALS algorithm (Mean Squared Error = 1.09519420585)
  - Dataset don’t perfectly match the requirement (User:Item=1:10)

- Product information were not fully made use of

Further Work

- Analyze ALS algorithm to understand why the top ten recommenders are centered around some users

- Take into account of group tag and category tag

- Compare the model performances on different “depth”