Master on Foundations of Data Science

Recommender Systems

Graph Based Models

Santi Seguí | 2017-2018

Modeling taste with Cassandra

Graph models for Neighborhood-Based Methods

- Sparsity of observed ratings causes a major problem in the computation of similarity in neighborhood-based methods.

- Graph-models can be used in order to **define similarity** in the neighborhood-based methods

  - using either structural transitivity or ranking techniques

- Provide a **structural representation** of the relationships **among** various **users** and/or **items**

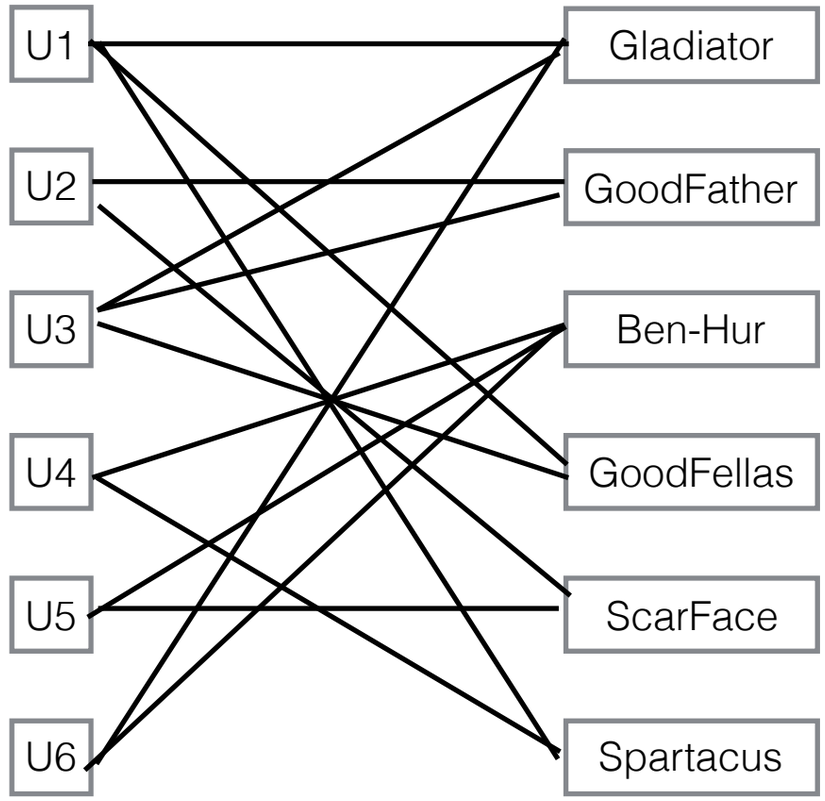
User-Item Graphs

- More effective than Pearson Correlation when dealing with very sparse datasets
- User-Item graph defined as an undirected and bipartite graph:

$$G = (N_u \cup N_i, A)$$

User-Item Graphs

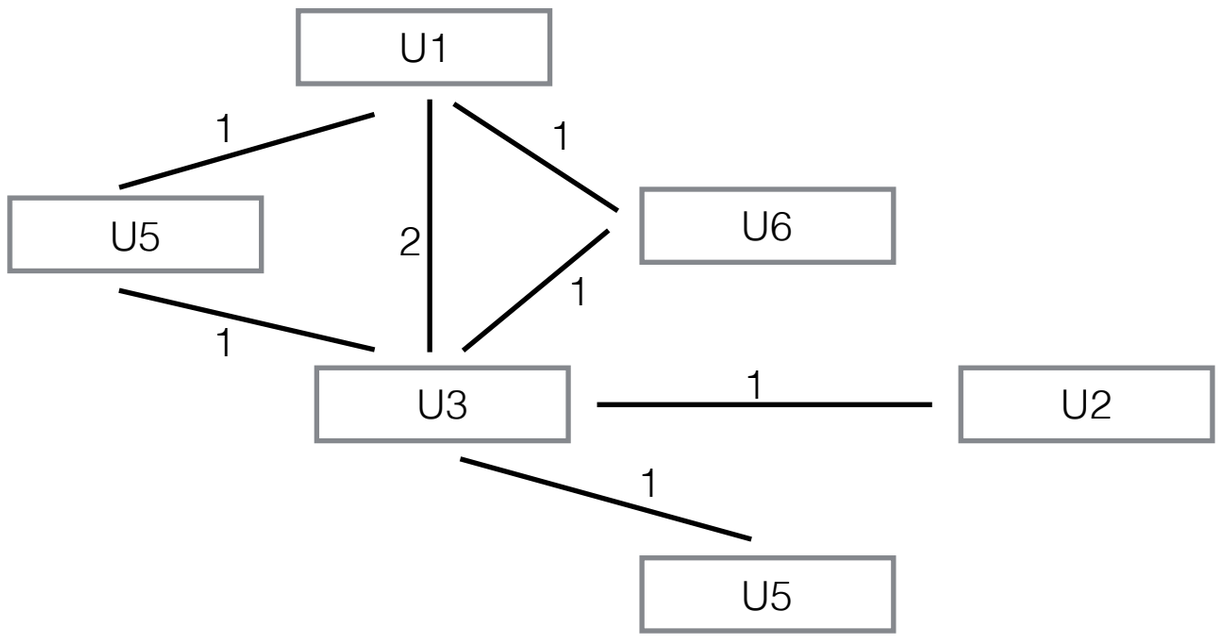
	Items					
	Gladiator	GoodFather	Ben-Hur	GoodFellas	ScarFace	Spartacus
U1	1			5		2
U2		5			4	
U3	5	3		1		
U4			3			4
U5				3	5	
U6	5		4			



User-User Graphs

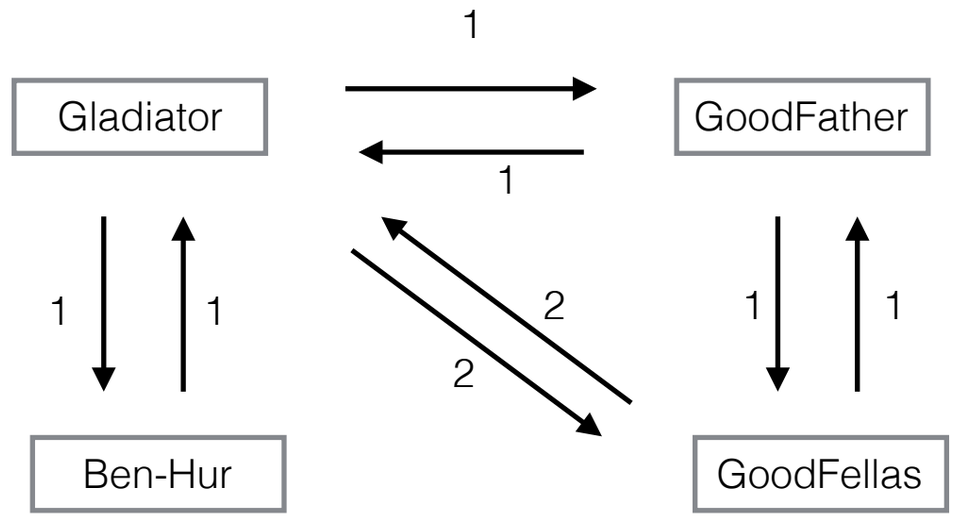
- User-user Graph based on 2-hop connectivity between users

	Gladiator	GoodFather	Ben-Hur	GoodFellas
U1	1			5
U2		5		
U3	5	3		1
U4			3	
U5				3
U6	5		4	

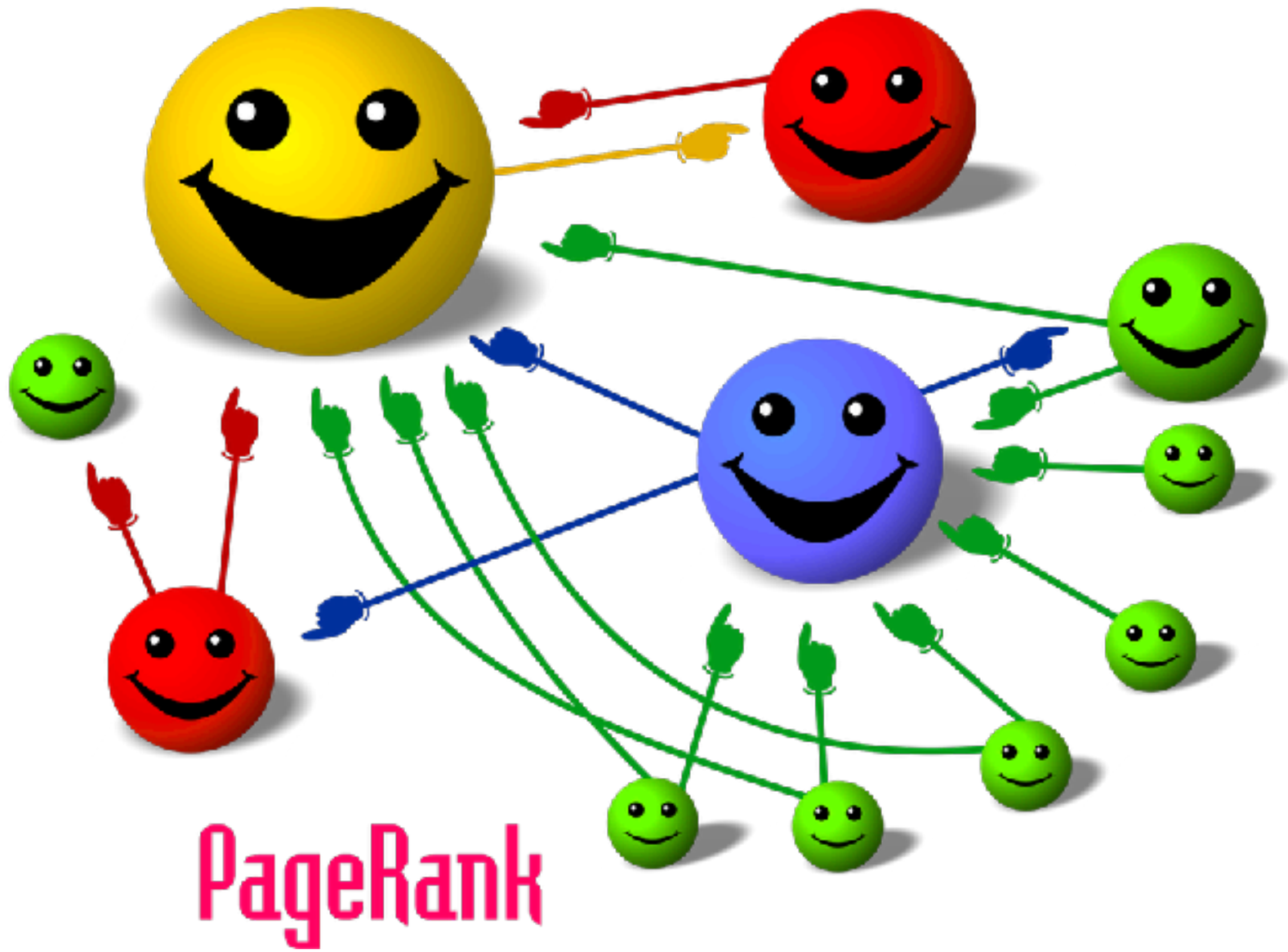


Item-Item Graphs

	Items			
	Gladiator	GoodFather	Ben-Hur	GoodFellas
U1	1			5
U2		5		
U3	5	3		1
U4			3	
U5				3
U6	5		4	



PageRank



PageRank

- The PageRank algorithm was first proposed in the context of Web search

- The PageRank algorithm generalizes the notion of citation-based ranking in a recursive way

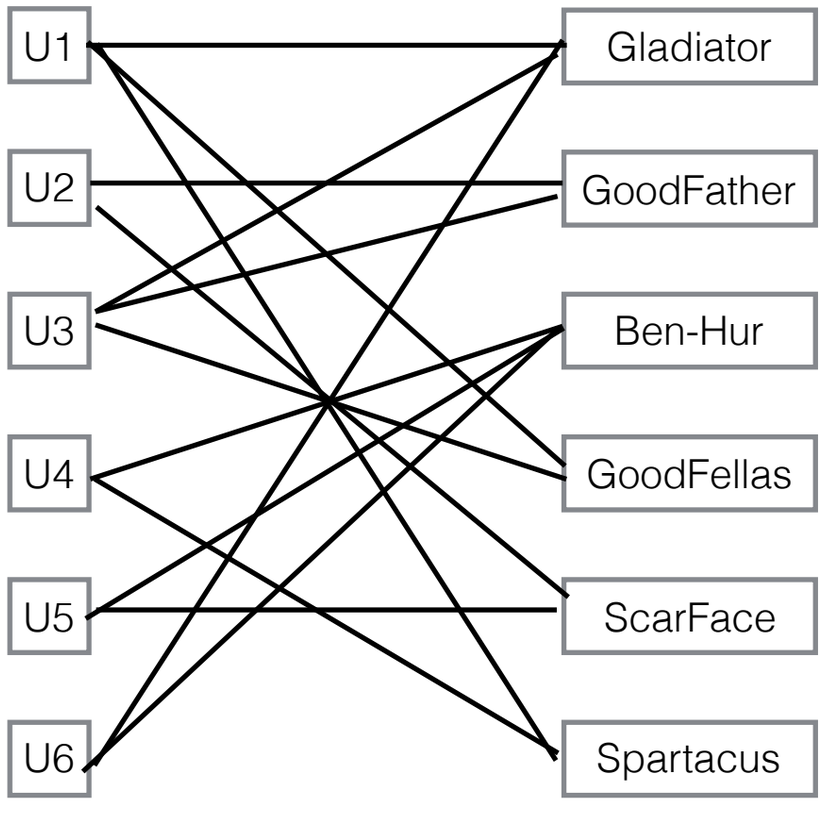
$$x' = (1 - \alpha)Ax + \alpha \frac{1}{n} S$$

Page Rank: How graph must be constructed?

Items

	Gladiator	GoodFather	Ben-Hur	GoodFellas	ScarFace	Spartacus
U1	1			5		2
U2		5			4	
U3	5	3		1		
U4			3			4
U5				3	5	
U6	5		4			

?	?	?	?	?



Page Rank: How graph must be constructed?

Items

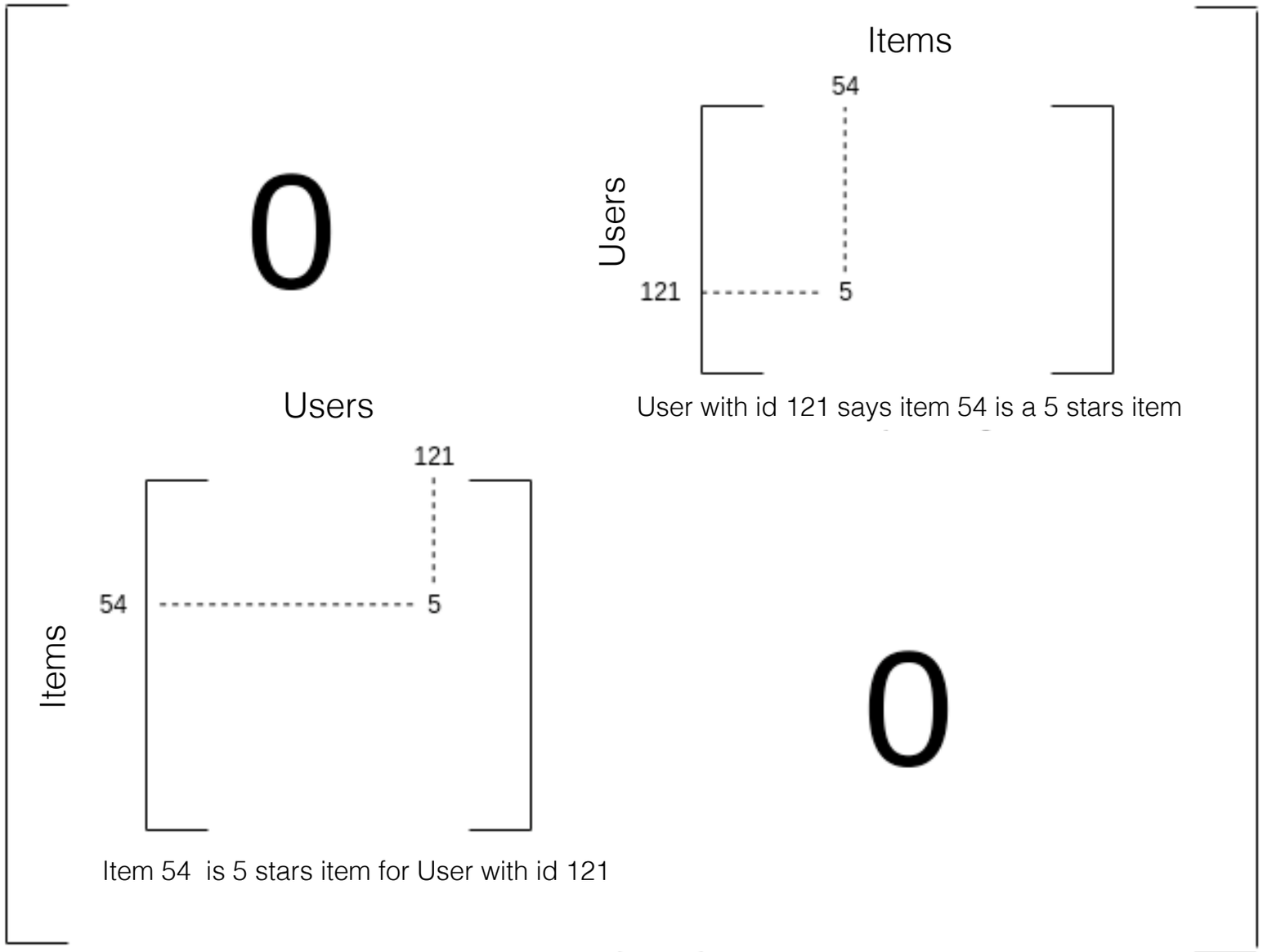
	U1	U2	U3	U5	U5	U6
U1	1			5		2
U2		5			4	
U3	5	3		1		
U4			3			4
U5				3	5	
U6	5		4			

User/User graph

?	?	?	?	?

User Weight?

Page Rank: Extended graph



PageRank is not directly a recommendation approach

it is Not personalized

Defining Neighborhoods

- The neighborhood of a user is defined by the set of users that are encountered frequently in a random walk starting at that user.

- How can we measure similarity between users/items using a graph?
  - Katz measure
  - **Personalized PageRank**
  - SimRank method

Personalized PageRank

- *PageRank* is an excellent mechanism to find popular nodes in terms of the linkage structure, however it does little for finding items that are well-matched to interest of specific users.

- The notion of *personalized PageRank* is designed to find **popular** nodes, which **are also similar to specific node** in the network

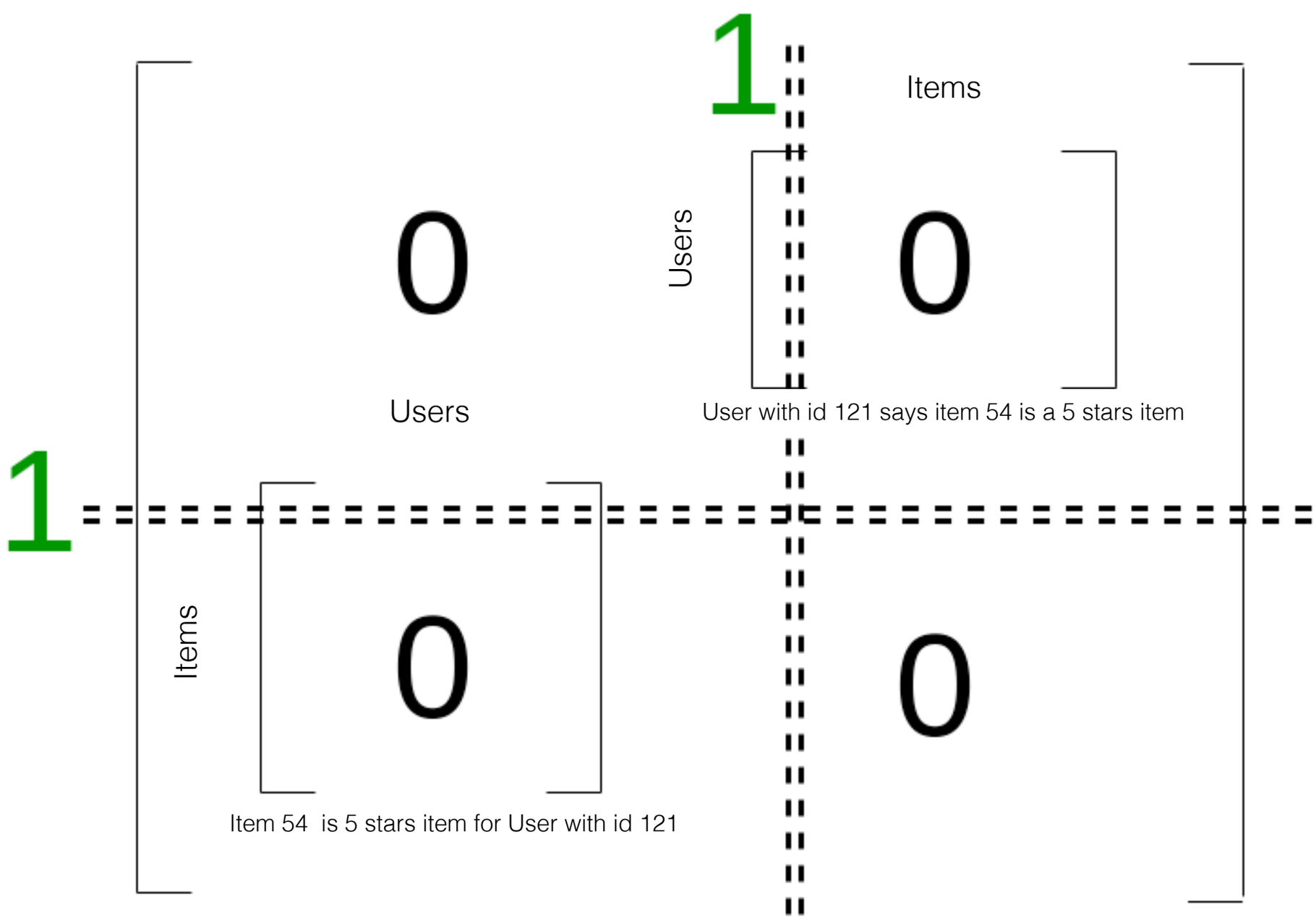
- A node receives an amount of rank from every node which points to it and in turn transfer an amount of its rank to the node it refers to.

Personalized PageRank

- Two main methods:
  - **Random walk** with restart at a particular item in order to determine the relevant neighborhoods
  - **ItemRank**. For each user *i*, a different PageRank restart vector is used.

Personalized PageRank

Random Walk



Personalized PageRank

- **ItemRank**. For each user *i*, a different PageRank restart vector is used.
- PageRank equations are specific to user *i* and one need to solve this system *m* times in order to determine the preferences of all users.

$$E(j) = \begin{cases} 1/n & \text{if } j \text{ in } I_u \\ 0 & \text{otherwise} \end{cases}$$

Task #3

- **Problem**: JOKES recommendations

- **Methods to implement:**
    - **Graph-Based recommender system**
    - Any other method you think will help you on the ranking
- **Evaluation**:
    - OFFLINE: MSE
- **Deadline**:
    - May 31th

InClass Prediction Competition

Jester Recommendations System

Jokes recommendations using jester dataset

21 days to go

Overview | Data | Kernels | Discussion | Leaderboard | Rules | Team | Host | My Submissions | **Submit Predictions**

Overview — Edit

Description

Evaluation

+ Add Page

- Jester dataset includes user ratings ranging from -10 to +10 for 100 jokes.
- NOTE: original dataset has been modified with noise and data perturbation
- RMSE will be used for evaluation.
- Once you submit a result, it will list you in the leaderboard based on the best score of your submissions.

TEAMS are allowed. From 1-3 persons

Leaderboard >

1 -
2 -
3 -
4 -
5 -
6 -

Kernels >

There are no kernels yet.

Be the first

0 discussion topics >

There are no topics yet.

Start one

Jester 5.0

Jokes for *your* sense of humor

First rate two jokes.

Q: If a person who speaks three languages is called "trilingual," and a person who speaks two languages is called "bilingual," what do you call a person who only speaks one language?

A: American!

Less Funny More Funny

Next