03

Data Characteristics of Recommendation Datasets

Recommendation data can be described by means of several properties, known as data characteristics.

The pioneering work from Adomavicius et al. [26] showed that recommenders' performance has a direct connection with these characteristics.

These results prompted further studies that focused on dataset characteristics to analyse their impact on various aspects of recommender systems [27, 28].

[26] Adomavicius & Zhang, ACM TMIS 2012, Impact of data characteristics on recommender systems performance
[27] Deldjoo et al., SIGIR 2020, How Dataset Characteristics Affect the Robustness of Collaborative Recommendation Models
[28] Deldjoo et al., IP&M 2021, Explaining recommender systems fairness and accuracy through the lens of data characteristics

By interpreting recommendation data as a bipartite graph, these characteristics can be extended to topological analyses.

A Novel Evaluation Perspective on GNNs-based Recommender Systems through the Topology of the User-Item Graph

Daniele Malitesta*†
Université Paris-Saclay
CentraleSupélec, Inria
Gif-sur-Yvette, France
daniele.malitesta@centralesupelec.fr

Claudio Pomo*
Politecnico di Bari
Bari, Italy
claudio.pomo@poliba.it

Vito Walter Anelli
Politecnico di Bari
Bari, Italy
vitowalter.anelli@poliba.it

Alberto Carlo Maria Mancino
Politecnico di Bari
Bari, Italy
alberto.mancino@poliba.it

Tommaso Di Noia
Politecnico di Bari
Bari, Italy
tommaso.dinoia@poliba.it

Eugenio Di Sciascio
Politecnico di Bari
Bari, Italy
eugenio.disciascio@poliba.it

In our study [29], we showed that these topological properties not only influence the final performance, but are also explicitly exploited by graph-based recommenders.

[29] Malitesta et al., RecSys 2024, A Novel Evaluation Perspective on GNNs-based Recommender Systems through the Topology of the User-Item Graph

Preliminaries

U: *user set*

I: *item set*

R: *number of ratings*

N_u: *user neighbourhood*

N_i: *item neighbourhood*

N^2_u: *user* 2_{nd} *order neighbourhood*

N^2_i: *item* 2_{nd} *order neighbourhood*

Space Size

Structural

Is one the 3 structural properties.

$$Space\ Size = |U| \times |I|$$

|I|

|U|

1		
	3	5
	5	4
2	1	
	4	4
1		5

User Item Ratio

Structural

Is one the 3 structural properties.

$$User\ Item\ Ratio = \frac{|U|}{|I|}$$

|U|

|I|

1		
	3	5
	5	4
2	1	
	4	4
1		5

Density / Sparsity

Structural

Is one the 3 structural properties.

$$Density = \frac{|R|}{|U| \times |I|}$$

|I|

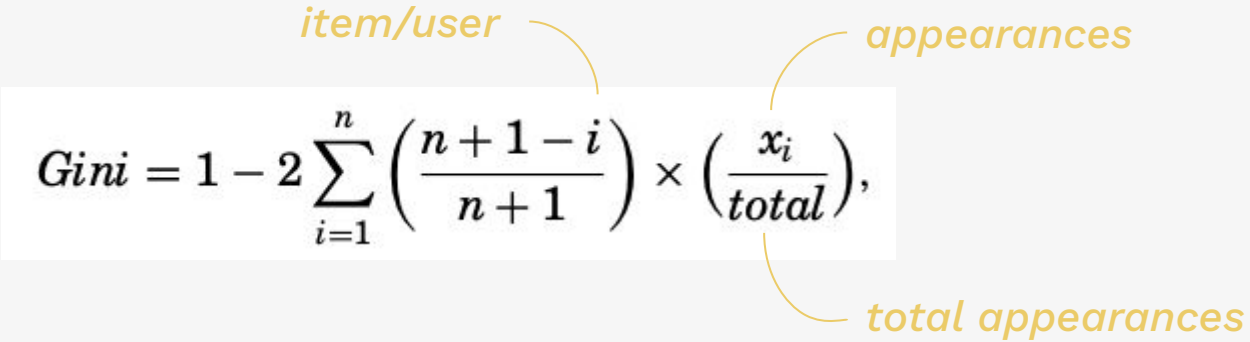
|U|

1		
	3	5
	5	4
2	1	
	4	4
1		5

Gini Item/User

Distributional

The Gini index measures inequality in the item/user frequency distribution.



A Gini index equal to 0 represents perfect equality (items equally popular / users with the same history length)

Variance

Value

The variance of ratings values is a measure of how frequent controversial items are in a dataset.

An item that receives **conflicting ratings** from different users is a controversial item.

A high variance in the ratings values is related to errors in the recommender's predictions.
The mean rating can also be used as a metric.

Node Degree

Topological

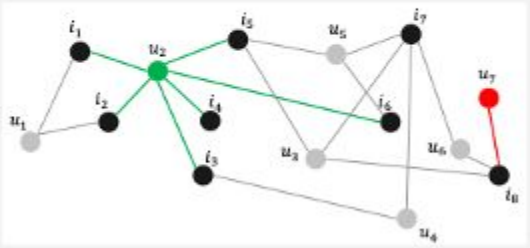
The node degree is a measure of the number of connections that a node has.

$$User\ node\ degree = |N_u|$$

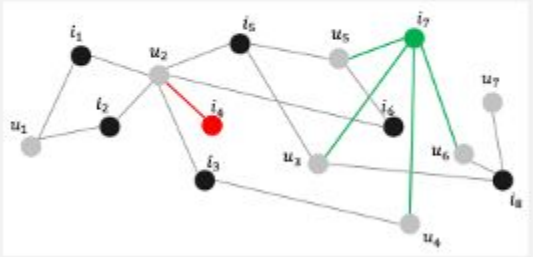
$$Item\ node\ degree = |N_i|$$

Node Degree in RecSys

A graphical representation of the node degree interpretation in the user-item graph. On the left, the user node degree as a measure of user activity. On the right, the item node degree distinguishes popular from niche items.



(a) active user *vs.* inactive user



(b) popular item vs. niche item

Clustering Coefficient

Topological

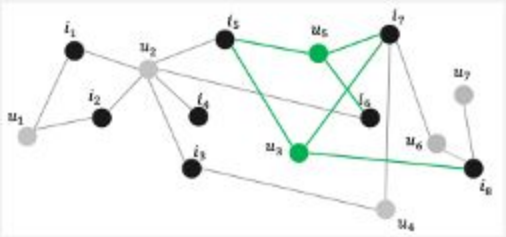
It measures the connectivity level of nodes of the same type within the graph

$$\gamma_{u,v} = \frac{|N_u \cap N_i|}{|N_u \cup N_i|}$$

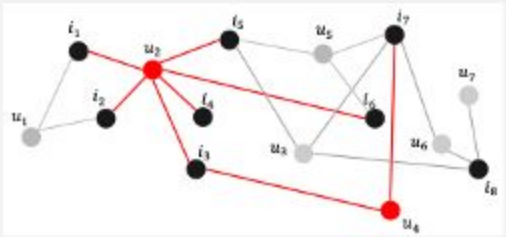
$$\gamma_u = \frac{\sum_{v \in N^2_u} \gamma_{u,v}}{|N^2_u|}$$

Clustering Coefficient in RecSys

Graphical representation of the node degree interpretation in the user-item graph. On the left, an example illustrates two similar users sharing the 67% of the interacted items. On the right, two neighboring users exhibit different preferences.



(a) similar user activity ($\gamma_{u_3,u_5} = 0.667$)



(b) different user activity ($\gamma_{u_2,u_4} = 0.167$)

Degree Assortativity

Topological

It measures the activity level similarity.

Fraction of edges linking nodes with two given degrees

Distinct degree values

$$\gamma_u = \frac{\sum_{d_h, d_k} d_h d_k \left(e_{d_h, d_k} - q_{d_h} q_{d_k}\right)}{std_q^2}$$

Probability of reaching a node from a node with the same degree

Std of the probability distribution

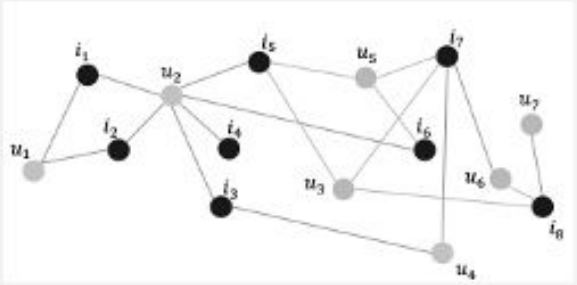
Degree Assortativity

Topological

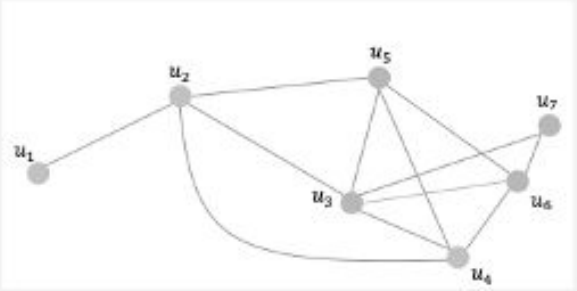
It measures the activity level similarity.

DEGREE ASSORTATIVITY IN RECSYS

Graphical representation of the degree assortativity interpretation in the user-item graph. In this example, user nodes are different in terms of node degree, thus resulting in a low value of degree assoratitvity.



(a) original user-item graph



(b) degree (**dis**)assortativity (ρ = -0.191)

Datasets Selection

Dataset characteristics are a useful tool to analyse and distinguish datasets even before carrying out the recommendation task.

Given the multitude of datasets available for offline evaluation, the literature still debates how a proper selection of datasets in an experiment should be carried out [30,31].

[30] Chin et al., WSDM '22, The Datasets Dilemma: How Much Do We Really Know About Recommendation Datasets?
[31] Vente et al., RecSys '25, APS Explorer: Navigating Algorithm Performance Spaces for Informed Dataset Selection