EVIA2025

ELIGE-IA

GOBIERNO DE ESPAÑA · MINISTERIO DE CIENCIA E INNOVACIÓN

Red Española de Investigación en **Sistemas de Recomendación**

Universidad de Jaén

Towards explainability and fairness in recommender systems

Luis Martínez López
martin@ujaen.es

Raciel Yera Toledo
ryera@ujaen.es

10th July 2025

OUTLINE

- **ELIGE IA**
- **RECOMMENDER SYSTEMS (RS)**
  - **GROUP RECOMMENDER SYSTEMS (GRS)**
- **Explainable AI (XAI)**
  - **Explainable Recommender Systems**
- **Fairness**
  - **Fair Recommender Systems**
- **CONCLUSIONS**

OUTLINE

- **ELIGE IA**
- RECOMMENDER SYSTEMS (RS)
  - GROUP RECOMMENDER SYSTEMS (GRS)
- Explainable AI (XAI)
  - RS and GRS
- Fairness
  - RS and GRS
- CONCLUSIONS

ELIGE IA

Red Temática Española de Investigación en Sistemas de Recomendación (ELIGE-IA)

- RED2022-134302-T

- **Universidad de Jaén (Coordinador)**
  - Luis Martínez/Raciel Yera
- **Universidad Politécnica de Madrid**
  - Jesús Bobadilla
- **Universidad de Barcelona**
  - María Sálamo
- **Universidad de Castilla La Mancha**
  - Jesús Serrano
- **Universidad de Santiago de Compostela**
  - Eduardo Sánchez
- **Universidad de Granada**
  - Carlos Porcel
- **Universidad Complutense**
  - María Belén Díaz
- **Universidad de Autónoma de Madrid**
  - Alejandro Bellogín
- **Universidad Rovira i Virgili**
  - Antonio Moreno
- **Universidad de Oviedo**
  - Antonio Bahamonde
- **Universidad Politécnica de Valencia**
  - Laura Sebastiá
- **Universidad Politécnica de Barcelona**
  - Miquel Sánchez-Marré

OUTLINE

- **RECOMMENDER SYSTEMS (RS)**
  - GROUP RECOMMENDER SYSTEMS (GRS)
- Explainable AI (XAI)
  - RS and GRS
- Fairness
  - RS and GRS
- IMPROVING Explainability and Fairness
  - Group Recommendation Systems
- CONCLUSIONS

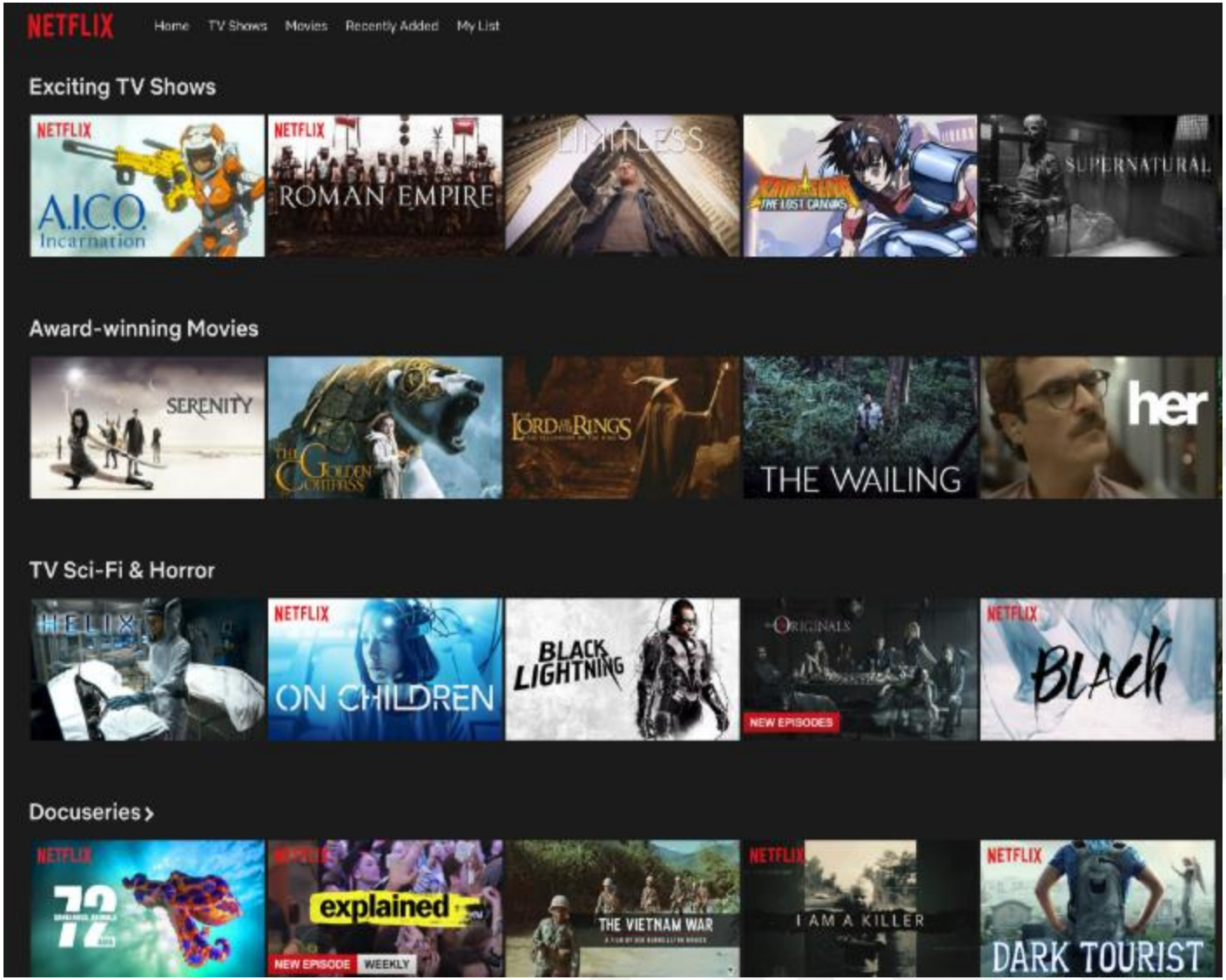
RECOMMENDER SYSTEMS

- Human Beings daily tasks
  - *What to wear?*
  - *What movie to rent?*
  - *What mobile to buy?*
  - The sizes of these decision domains are frequently big
    - Internet: ***Massive***
      - Netflix has over 50,000 movies
      - Amazon.com has over 32 million books in the book store

RECOMMENDER SYSTEMS

- Internet
  - Information Overload
  - Explore and Filter out
    - Irrelevant Information
    - Preferences and needs

} **Support**


NETFLIX
Home
TV Shows
Movies
Recently Added
My List
Exciting TV Shows
A.I.C.O. Incarnation
ROMAN EMPIRE
LIMITLESS
SUPERNATURAL
Award-winning Movies
SERENITY
LORD OF THE RINGS
THE WAILING
her
TV Sci-Fi & Horror
HELIX
ON CHILDREN
BLACK LIGHTNING
ORIGINALS
NEW EPISODES
BLACK
Docuseries
explained
NEW EPISODE
WEEKLY
THE VIETNAM WAR
I AM A KILLER
DARK TOURIST

19 min searching

at least once a week

300 million viewers

5 billion hours wasted per year

RECOMMENDER SYSTEMS

- Internet
  - Information Overload
  - Explore and Filter out
    - Irrelevant Information
    - Preferences and needs

  } **Support**

- Personalisation ⇨ **Information and ecommerce ecosystem**
  - Tailor content
  - Services

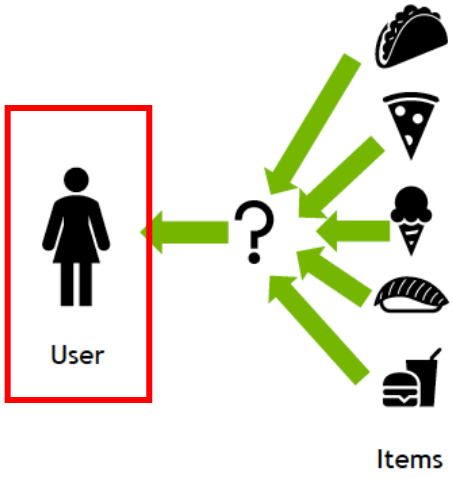
  } **INDIVIDUALS** {
  - Preferences
  - Behaviour

RECOMMENDER SYSTEMS

OVERVIEW

Recommender Systems (RS) typically apply techniques and methodologies from other areas – such as Human Computer Interaction (HCI) or Information Retrieval (IR). However, most of these systems bear in their core an algorithm that can be understood as a particular instance of a Artificial Intelligence technique

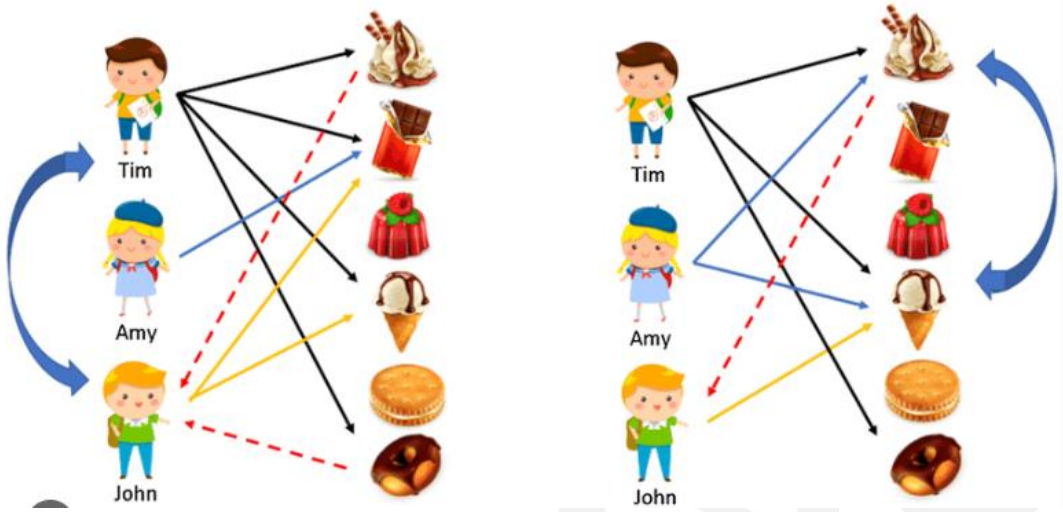
DEFINITION



(Burke, 2002) ***"… any system that produces individualized recommendations as output or has the effect of guiding the user in a personalized way to interesting or useful objects in a large space of possible options."***

RECOMMENDER SYSTEMS

- **Main research stream**
  - Most of RS recommend items to individual users
  - Enhance recommendation
  - Effectiveness for users based on their past interactions
  - Evaluated by typical ranking-based metrics

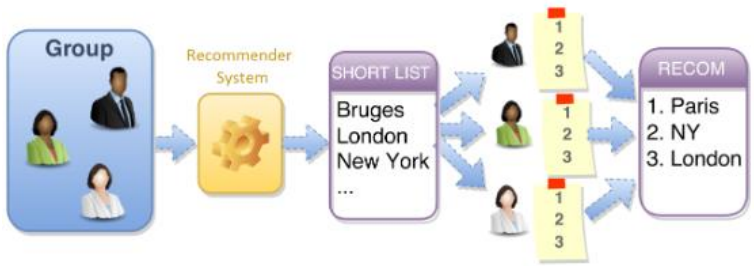


OUTLINE

- **RECOMMENDER SYSTEMS (RS)**
  - **GROUP RECOMMENDER SYSTEMS (GRS)**
- **Explainable AI (XAI)**
  - **RS and GRS**
- **Fairness**
  - **RS and GRS**
- **IMPROVING Explainability and Fairness**
  - **Group Recommendation Systems**
- **CONCLUSIONS**

GROUP RECOMMENDER SYSTEMS

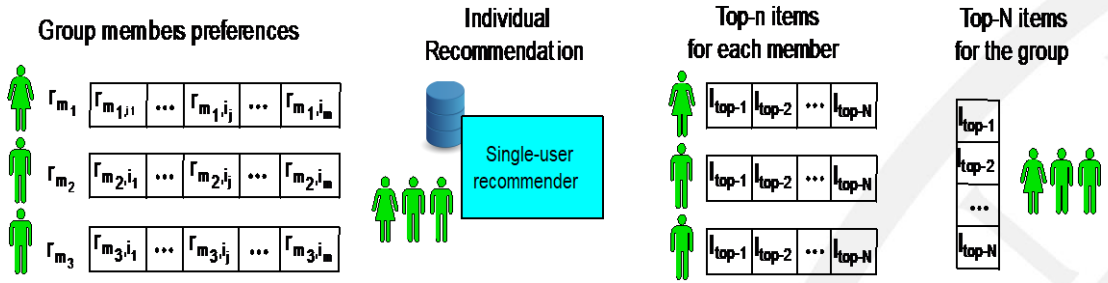
- **Recommending to groups**
  - Most of RS recommend items to individual users
  - One step further
    - Situations in which recommend to a group would be good
      - ***Social products***
        - Restaurants, POIs, social-nets content
        - Television programs
        - Songs to listen
  - Group recommendation more complex than individual



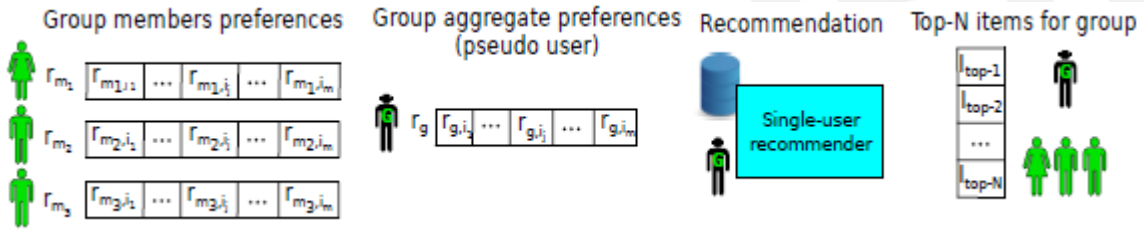
J. Masthoff, Group Recommender Systems: Aggregation, Satisfaction and Group Attributes, in: F. Ricci, L. Rokach, B. Shapira (Eds.), Recommender Systems Handbook, Springer US. 2015, pp. 743–776. ISBN 978-1-4899-7636-9.

GROUP RECOMMENDER SYSTEMS

- **Group recommendation Models**
  - *Main Techniques*:
  - *Aggregation* of individual **recommendations**



  - Users' *Profile Aggregation*



OUTLINE

- RECOMMENDER SYSTEMS (RS)
  - GROUP RECOMMENDER SYSTEMS (GRS)
- **Explainable AI (XAI)**
  - RS and GRS
- Fairness
  - RS and GRS
- IMPROVING Explainability and Fairness
  - Group Recommendation Systems
- CONCLUSIONS

EXPLAINABLE AI

- **AI based systems**
  - Scrutinized by governments and administrations
  - Need for decision-making robust, transparent, and accountable
  - EU has a greater significance
    - Ethics Guidelines for a Trustworthy Artificial Intelligence (2019)
    - Recent Provisional agreement on the artificial intelligence act (2021-…)

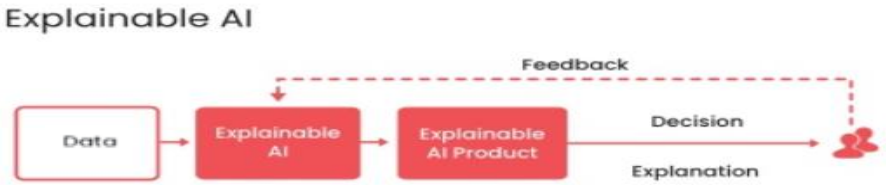
AI systems should adhere *to the* ***ethical principles*** *of respect for human autonomy, prevention of harm*

- *Explainability*: better understanding of how AI thinks and provides a solution
- *Fairness*: ensure that individuals and groups are free from unfair bias and discrimination in AI

EXPLAINABLE AI

- **IA based systems**
  - Why XAI is needed?

  - Misperception with AI black box
    - Why did the AI system do that and not other thing?
    - When do the AI system succeed or fail?



  - Clear and Transparent Solutions
    - It is easy to understand why and why not?
    - It is easy to understand when succeed or fail
    - If solution is understandable the system is trustworthy

OUTLINE

- RECOMMENDER SYSTEMS (RS)
  - GROUP RECOMMENDER SYSTEMS (GRS)
- **Explainable AI (XAI)**
  - **RS and GRS**
- Fairness
  - RS and GRS
- IMPROVING Explainability and Fairness
  - Group Recommendation Systems
- CONCLUSIONS

EXPLAINABLE RECOMMENDER SYTEMS

- **Explainability: AI Tools vs RECOMMENDER SYSTEMS**
  - Computer vision: Image Classification
    - Models are built to accurately predict the ***labels*** like: cat or dog
    - These labels are objective facts
    - Do not change with human feelings
  - Other domains
    - The tasks are more subjective
    - There is no rigid ground truth
    - The target is to improve the *utilities* of certain stakeholders
  - **Recommender system** is a typical subjective AI task
    - Utilities not only include accuracy, but also other aspects
    - ***Explainability*** is a significant and widely studied one.

EXPLAINABLE RECOMMENDER SYTEMS

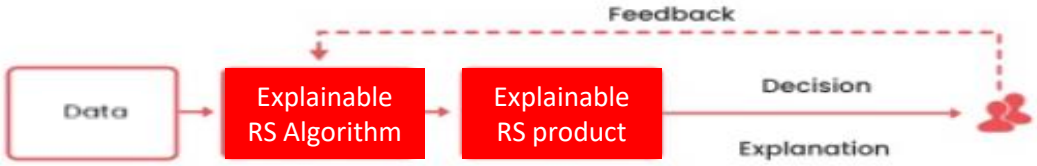
- **RECOMMENDER SYSTEMS**
  - Why XRS are needed?
    - Lack explainability of RS exists
      - The output of the recommendation systems
      - The mechanism of the recommendation model
        - The recommendation algorithm
      - This lack of explainability of recommendation algorithm
        - Leads to problems
          - Users do not know why specific results are provided?
          - System may be is less effective in persuading users
          - Decrease system's trustworthiness

EXPLAINABLE RECOMMENDER SYTEMS

- **RECOMMENDER SYSTEMS**
  - XRS aims at:
    - **Model Validation:** Avoid biases, overfitting or detect issues in the training data, adhere to ethical/legal requirements
    - **Knowledge Discovery:** explanations provide feedback to users that can result in new insights by revealing hidden underlying correlations/patterns
    - **Trust:** explanations might convince users to adopt the IA based technology

Black Box RS

Explainable RS



EXPLAINABLE RECOMMENDER SYTEMS

- **RECOMMENDER SYSTEMS**
  - Explainable recommendation solves:
    - **Why** the items are recommended?
    - With Explanations the RS:
      - Improving recommendation persuasiveness
      - User satisfaction
      - System transparency

EXPLAINABLE RECOMMENDER SYTEMS

- **RECOMMENDER SYSTEMS**
  - Why XRS are needed?
    - Nowadays RS not only for seeking information but also
      - Complicated decisión-making
        - Medical workers need comprehensive document recommendations
        - Retrieval to make diagnosis
      - Explanaibility of these results are extremly needed

Zhang, Y., & Chen, X. (2020). Explainable recommendation: A survey and new perspectives. Foundations and Trends in Information Retrieval, 14(1), 1-101.

EXPLAINABLE RECOMMENDER SYTEMS

Intrinsic models

Focused on explaining the model behavior

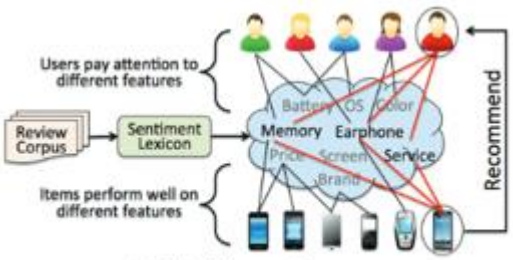
Post-hoc models

Focused on explaining the model output

Explanation interfaces for recommender systems

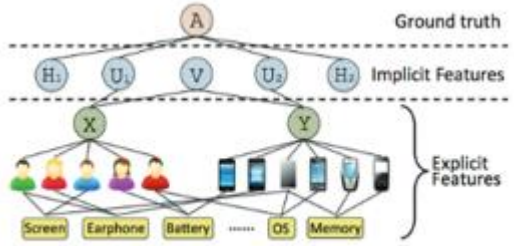
Zhang, Y., & Chen, X. (2020). Explainable recommendation: A survey and new perspectives. Foundations and Trends in Information Retrieval, 14(1), 1-101.

EXPLAINABLE RECOMMENDER SYTEMS



(a) Model overview

(b) Features and sentiment



(c) Multi-matrix factorization

You might be interested in [feature], on which this product performs well.

You might be interested in [feature], on which this product performs poorly.

(d) Recommendation explanations

- **INTRINSIC EXPLICIT MODEL**
  - **Idea:** Recommend ítems that perform well on user's favorite features
  - **Review:** it is transformed into a set of product features with user's sentiment on features
  - **User-attention feature and ítem quality matrixes:** to predict rating matrix
  - **Explicit product features:** Used to generate explanations

Zhang, Y., G. Lai, M. Zhang, Y. Zhang, Y. Liu, and S. Ma (2014a). Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval. ACM. 83–92.

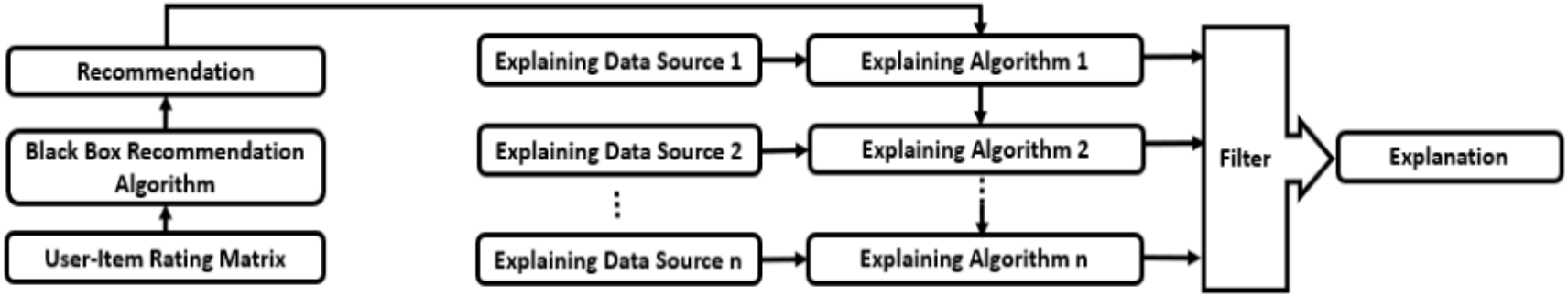
EXPLAINABLE RECOMMENDER SYTEMS

- POST HOC MODELS
  - Focused on coupling any main recommendation method with a white-box explainable method that facilitates the explanation of the main approach
  - These models are built on the assumption that an explanation that makes sense to the user, even if it is not the exact reason that the recommendation was indeed issued, is acceptable to users and may have a benefitial effect for the RS

EXPLAINABLE RECOMMENDER SYTEMS

- POST HOC MODELS
  - These models do not make necessarily transparent the recommendation algorithm because the explanation cannot show the exact reason for issuing its recommendations

Anatomy of a post-hoc explanation model in RS



Shmaryahu, D., Shani, G., & Shapira, B. (2020, January). Post-hoc Explanations for Complex Model Recommendations using Simple Methods. In IntRS@ RecSys (pp. 26-36).

EXPLAINABLE RECOMMENDER SYTEMS

- **Post-hoc explanation contextualized to specific scenarios**
  - Post-hoc support by simple models

Shmaryahu, D., Shani, G., & Shapira, B. (2020, January). Post-hoc Explanations for Complex Model Recommendations using Simple Methods. In IntRS@ RecSys (pp. 26-36).

  - Explanation Mining

Peake, G., & Wang, J. (2018, July). Explanation mining: Post hoc interpretability of latent factor models for recommendation systems. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2060-2069).

  - LIME

Nóbrega, C., & Marinho, L. (2019, April). Towards explaining recommendations through local surrogate models. In Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing (pp. 1671-1678).

EXPLAINABLE RECOMMENDER SYTEMS

- Post-hoc supported by simple models (for recipe recommendation)

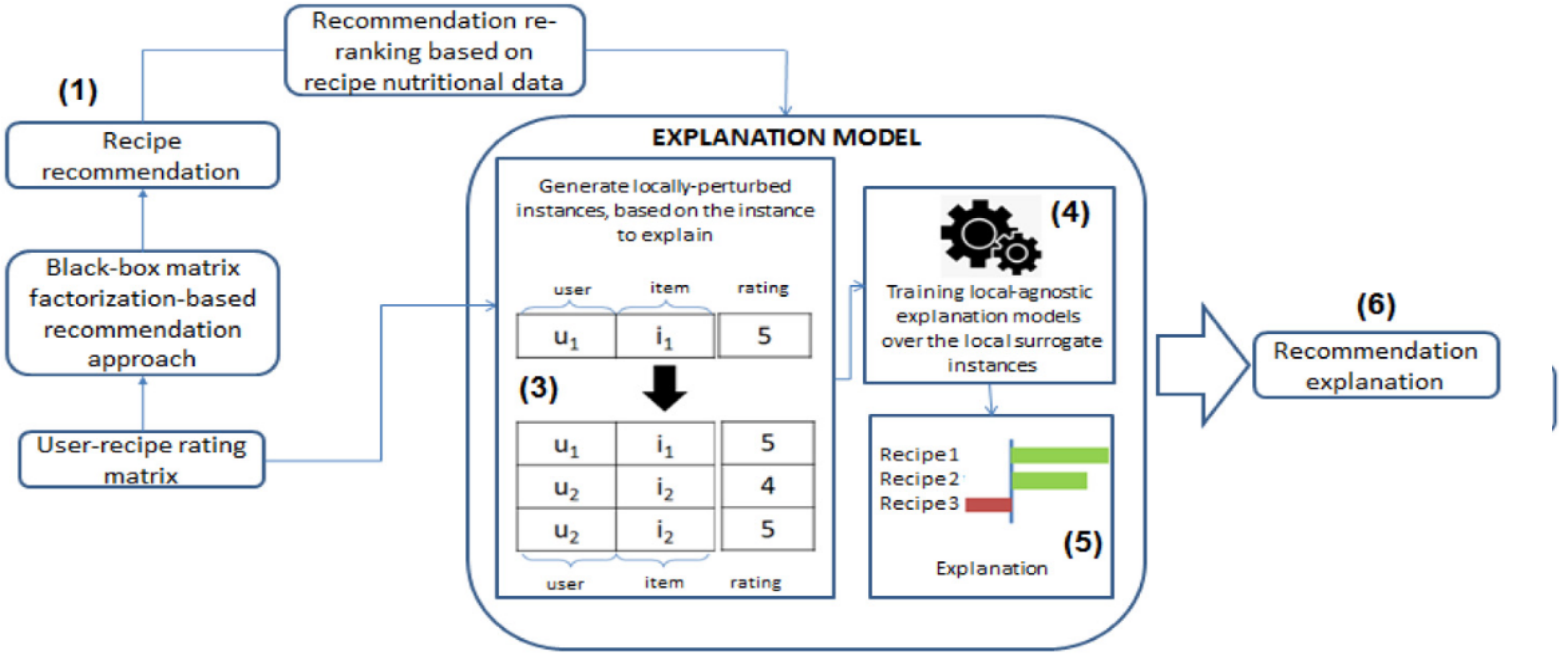


Fig. 5. Locally interpretable model-agnostic explanation model.

Yera, R., Alzahrani, A. A., & Martínez, L. (2022). Exploring post-hoc agnostic models for explainable cooking recipe recommendations. Knowledge-Based Systems, 251, 109216.

EXPLAINABLE RECOMMENDER SYTEMS

- Results of explanaibility models for nutrition
  - Explaination mining models
    - Higher Fidelity than simple-to-explain models
  - Important overlapping among methods
  - In some cases one method can act as input of another
  - Including nutrition-aware criteria do not provide significant improvements of recommendations

Yera, R., Alzahrani, A. A., & Martínez, L. (2022). Exploring post-hoc agnostic models for explainable cooking recipe recommendations. Knowledge-Based Systems, 251, 109216.

Explanations for groups

EXPLAINABLE GROUP RECOMMENDER SYTEMS

Explanations built over the social choice-based aggregation strategies

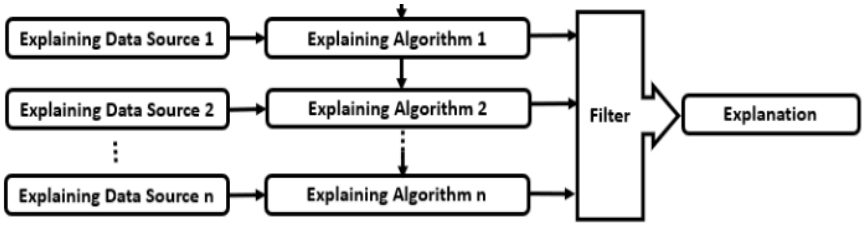
Strategy	Basic explanation	Detailed explanation
Additive	"i_k has been recommended to the group since it achieves the highest total rating."	"i_k has been recommended to the group since it achieves the highest total rating (as the sum of the ratings of all members for i_k is r which is higher than other items)."
Least Misery	"i_k has been recommended to the group since no group members has a real problem with it."	"i_k has been recommended to the group since no group members has a real problem with it (as $u_{j_1}, u_{j_2}, \ldots$ and $u_{j_{\bar{n}}}$ gave it a rating of r which is the highest rating among the lowest ratings per item)."
Most pleasure	"i_k has been recommended to the group since most group members like it."	"i_k has been recommended to the group since most group members like it (as $\bar{n}$ out of n group members gave it a high rating)."

Barile, F., Draws, T., Inel, O., Rieger, A., Najafian, S., Ebrahimi Fard, A., ... & Tintarev, N. (2024). Evaluating explainable social choice-based aggregation strategies for group recommendation. User Modeling and User-Adapted Interaction, 34, 1-58.

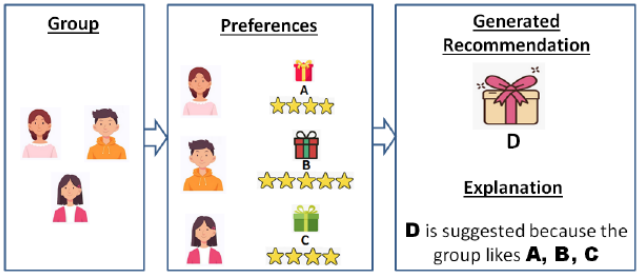
EXPLAINABLE GROUP RECOMMENDER SYTEMS

NOVEL PROPOSAL

Post-hoc explanations



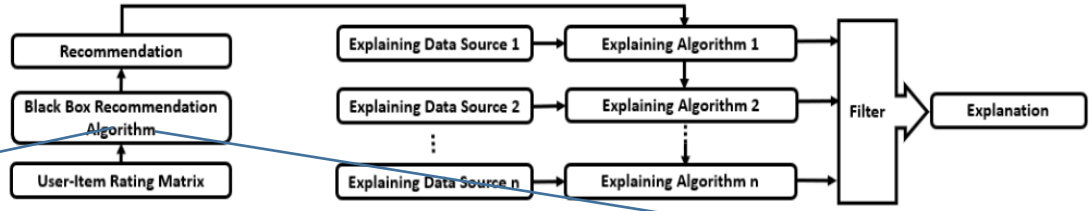
Group recommender systems



Yera, R., & Martínez L. (2024). Towards post-hoc explanation approaches in group recommendation. In *Spanish Conference of Artificial Intelligence*. A Coruña, Spain.

EXPLAINABLE GROUP RECOMMENDER SYTEMS

FIRST PROPOSAL



"Type I explanations: The item D is recommended because at least K members of the group preferred items A, B, and C".

"Type II Explanations: The item D is recommended because all the members of the group preferred at least A, B, or C".

$$I_{u_1} = \{i_1^{u_1}, i_2^{u_1}, \ldots\}$$

$$I_{u_2} = \{i_1^{u_2}, i_2^{u_2}, \ldots\}$$

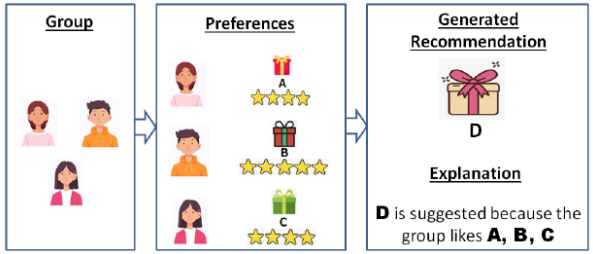
$$I = I_u$$

(at least for K users in the group)

Rule:

(A,B,C) -> D

$$I \subseteq \bigcup_{u \in G} I_u \qquad \forall u \in G \qquad I \cap I_u \neq \varnothing$$



EXPLAINABLE GROUP RECOMMENDER SYTEMS

Dataset: Movielens 100K

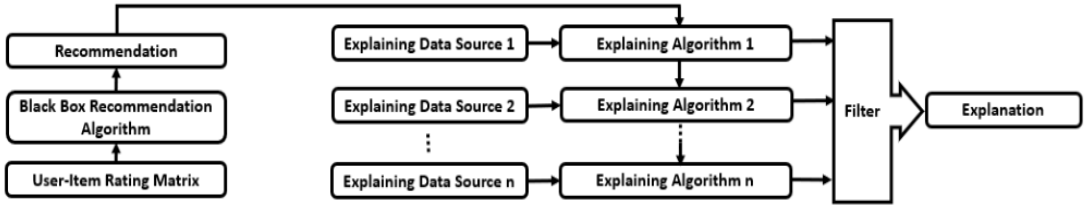
$$Fidelity(M) = \frac{|set\ of\ recommended\ items \cap set\ of\ justifiable\ recommendations|}{|set\ of\ recommended\ items|}$$

TABLE I
FIDELITY VALUES

	n=1	n=3	n=4
Type I Explanations			
Group size 5	0.78	0.71	0.23
Group size 6	0.71	0.66	0.45
Type II Explanations			
Group size 5	0.62		
Group size 6	0.59		

EXPLAINABLE GROUP RECOMMENDER SYTEMS

SECOND PROPOSAL



Explanations: The item D is recommended based on the feature values F1=v1 and F2=v2.

Rule:

(F1=v1,F2=v2) -> D

Also introduces counterfactual explanations, focused on answering:

Which are the minimum changes in feature values for becoming the item into *not recommended*?

Yera, R., & Martínez, L. (2024). LORE4GroupRS: Explaining Group Recommendations Supported by a Local Rule-Based Approach. In *International Conference on Intelligent Data Engineering and Automated Learning* (pp. 300-312). Cham: Springer Nature Switzerland.

Methodology

1) Individual and Group Recommendation Generation

2) Explanations Generation for the Individual Recommendations

3) Explanation Generation for Groups

1) Individual and Group Recommendation Generation

$U = \{u_1, u_2, u_3, ..., u_n\}$ Set of users

$I = \{i_1, i_2, i_3, ..., i_n\}$ Set of items

$R = \{r_{ui}\}$ Set of ratings

$p_{ui} = M(u, i)$ Prediction of unknown ratings

$Rec_{u_k} = \{i\}, \quad \forall_i \; p_{u_k,i} \geq pref_{th}$ Top k items recommendation

The individual recommendations generated for each group member are then combined using different aggregation approaches, for obtaining the group recommendations.

2) Explanations Generation for the Individual Recommendations (1)

Content-based item profile: $prof_i = \{(a_k = v_j^{a_k})\}, \forall_{a_k} a_k \in A$

Set of attributes A

Similarity between item profiles: $sim(prof_{i1}, prof_{i2}) = \frac{|prof_{i1} \cap prof_{i2}|}{|prof_{i1} \cup prof_{i2}|}$

Individual recommendation explanation: $e =< r, \phi >$

$r = \{(a_k = v_j^{a_k})\} \rightarrow i$ Factual explanation

$\phi = \{\{(a_{k_1} = v_j^{a_{k_1}})\}, \{(a_{k_2} = v_j^{a_{k_2}})\}, ...\}$ Counterfactual

2) Explanations Generation for the Individual Recommendations (2)

Goal: Calculating the set of individual recommendation explanations E_i, for each item *i* in the set of recommended items,

Approach:

1) Item's Neighborhood Calculation.

2) Extraction of Local Rule-Based Recommendation Explanations

2) Explanations Generation for the Individual Recommendations (3)

1) Item's Neighborhood Calculation.

$$I_i = \{j\}, \forall_j sim(prof_i, prof_j) \geq sim_{th}$$

Based on I_i, it is built a set of instances Z that will be the input for the next step, composed of:

1. A set of feature characteristics, represented by all the terms in the profile $prof_j$.
2. A decision $d(j)$ associated to the current item, that could be their recommendation j, or not recommendation $\hat{j}$.

2) Explanations Generation for the Individual Recommendations (4)

2) Extraction of Local Rule-Based Recommendation Explanations

It is processed the set of instances *Z* composed of items *j* which were detected in the neighborhood of the recommended *i* to be explained.

For this set of instances, it is constructed a decision tree built over the qualitative attribute values of the items for composing the branches, and the decision of *Recommendation* (d(j)=1) or *NoRecommendation* (d(j)=0) for each *j* as leaves of the tree.

Factual explanation:

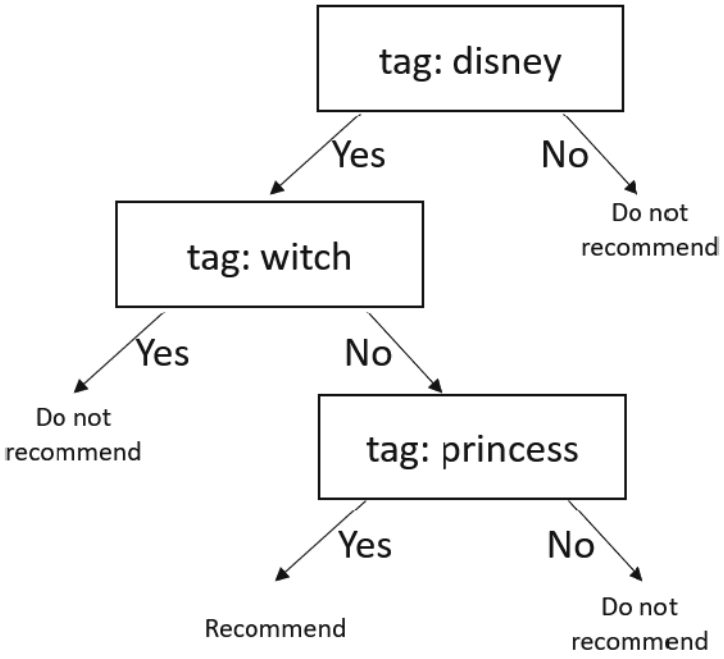
Joining of verified conditions located at a path of the tree from the root to the leaf satisfying d(j)=1

Counterfactual explanation:

Paths from the root to the leaves, which have in the leave the decision of *NoRecommendation*, having minimal difference with the factual one

2) Explanations Generation for the Individual Recommendations (4)

Scenario:



r = (*disney* : *yes*,*witch* : *no*, *princess* : *yes*) → *recommend*

φ = (((*disney* : *yes*,*witch* : *no*, *princess* : *no*) → *donotrecomm*), ((*disney* : *yes*,*witch* : *yes*) → *donotrecommend*)).

3) Explanation Generation for Groups

Approach

1) Perform the union of all the split conditions at the r components associated to each group member, into a group factual explanation.

2) Remove contradictory conditions.

3) Perform the union of all the counterfactuals.

Experiments (1)

Movielens 100K dataset

Features: Movies' tags

Individual recommendation approach: Koren's SVD

Evaluation criteria:

$$ModelFidelity(M) = \frac{|recommended\ and\ explainable\ items(M)|}{|recommendeditems(M)|}$$

$$\text{Feature Coverage Ratio (FCR)} = \frac{N_a}{|A|}$$

Experiments (2)

Parameters:

-Amounts of used tags: *m=10 and m=40*

-Item similarity threshold for building the local neighborhood: *sim_th = [0; 0.3]*

-Preference threshold for choosing an ítem as recommended: *pref_th=[3; 4.2]*

-Stopping criteria for the decision tree building: *stop_th=0.2*.

Main findings (1). Individual recommendations.

Table 1. Performance of the approach for the individual recommendation scenario. $m = 40$, $pref_{th} = 3.6$.

sim_{th}	0	0.05	0.1	0.15	0.2	0.25	0.3
Model Fidelity	0.7483	**0.7925**	0.7174	0.6755	0.6667	0.6225	0.553
Feature Coverage Ratio	**0.725**	0.65	0.675	0.525	0.55	0.275	0.175

Table 3. Performance of the approach for the individual recommendation scenario. $m = 10$, $pref_{th} = 3.6$.

sim_{th}	0	0.05	0.1	0.15	0.2	0.25	0.3
Model Fidelity	**0.7842**	0.708	0.708	0.708	0.7118	0.7209	0.7183
Feature Coverage Ratio	**0.9**	0.7	0.7	0.7	0.7	0.7	0.7

- The best performance in terms of model fidelity, was obtained for sim_th=0.05.
- The overall results for model fidelity and feature coverage ratio are not correlated.
- The use of les tags does not necessary imply a lower fidelity.

Main findings (2). Group recommendations.

Table 5. Performance of the approach for the group recommendation scenario. $m = 40$, $pref_{th} = 3.6$.

sim_{th}	0	0.05	0.1	0.15	0.2	0.25	0.3
Model Fidelity	**0.2097**	0.1129	0.1290	0.08065	0.0483	0.0968	0.0322
Feature Coverage Ratio	**0.325**	0.2	0.2	0.15	0.125	0.05	0.05

Table 7. Performance of the approach for the group recommendation scenario. $m = 10$, $pref_{th} = 3.6$.

sim_{th}	0	0.05	0.1	0.15	0.2	0.25	0.3
Model Fidelity	**0.2**	0.1538	0.1538	0.1538	0.1538	0.1231	0.1231
Feature Coverage Ratio	**0.7**	0.5	0.5	0.5	0.4	0.5	0.3

- Group recommendation leads to lower values of model fidelity and FCR, in relation to individual.
- Here the best performance in terms of model fidelity, was obtained for sim_th=0.

EXPLAINABLE RECOMMENDER SYTEMS

- **RECOMMENDER SYSTEMS**
  - How to evaluate explainability? What strategies are better?
    - Evaluating recommendation explanations is much more difficult
      - The ground truth is hard to obtain
      - Human feelings are not easy to approximate
    - To alleviate these difficulties
      - A lot of promising evaluation strategies
      - But lacks a systematic comparison between them

EXPLAINABLE RECOMMENDER SYTEMS

- **RECOMMENDER SYSTEMS**

  - Explanations should be evaluated from different views

Evaluation perspective	Evaluation problem	Serving target
Effectiveness	Whether the explanations are useful for the users to make more accurate/faster decisions?	Users
Transparency	Whether the explanations can reveal the internal working principles of the recommender models?	Model designers
Persuasiveness	Whether the explanations can increase the click/purchase rate of the users on the items?	Providers
Scrutability	Whether the explanations can exactly correspond to the recommendation results?	Model designers

EXPLAINABLE RECOMMENDER SYTEMS

- **RECOMMENDER SYSTEMS**
  - Explanations evaluation methods

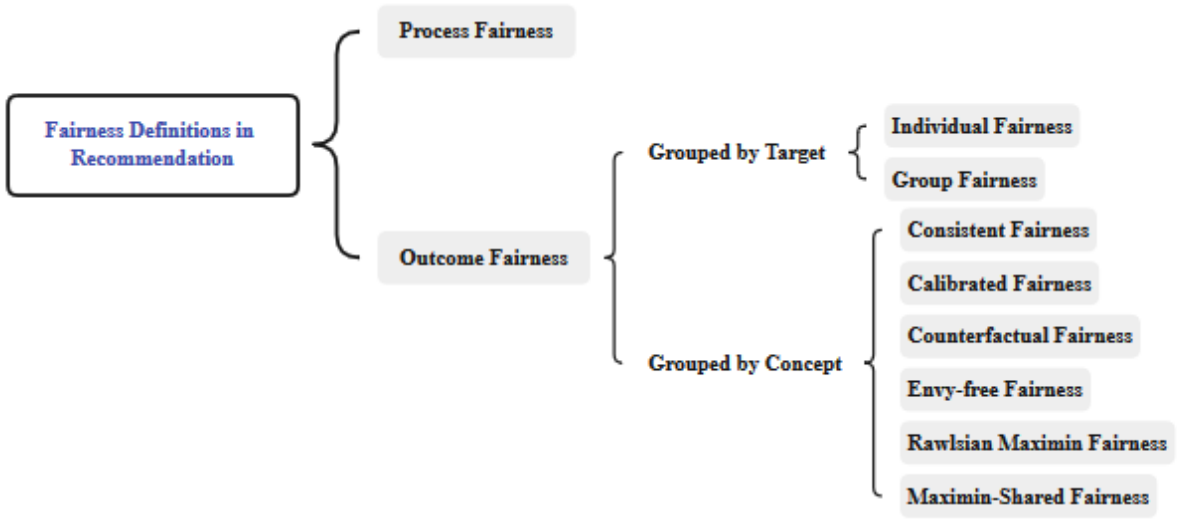
Method	Advantages	Disadvantages	Typical Evaluation Perspectives
Case Studies	Intuitive and human-understandable	Subjective, biased, not scalable	Effectiveness, Transparency
Quantitative Metrics	Objective, efficient, benchmarkable	May misalign with actual user utility	Effectiveness, Scrutability
Crowdsourcing	Real user feedback, subjective fidelity	High cost, limited scalability	All: Effectiveness, Persuasiveness, Transparency, Scrutability
Online Experiments	High external validity	Expensive, operationally disruptive	Effectiveness, Persuasiveness

OUTLINE

- RECOMMENDER SYSTEMS (RS)
  - GROUP RECOMMENDER SYSTEMS (GRS)
- Explainable AI (XAI)
  - RS and GRS
- **Fairness**
  - RS and GRS
- IMPROVING Explainability and Fairness
  - Group Recommendation Systems
- CONCLUSIONS

FAIRNESS

- How the system treats people, or groups of people
  - In a “unfair” way by some moral, legal, or ethical standard.
- It is a *human value* discussed in many disciplines, now in AI.
  - It is a developing area in Recommender Systems
  - Countless definitions of fairness have been proposed
  - Primary notions of fairness in recent literature



Ekstrand, M. D., Das, A., Burke, R., & Diaz, F. (2022). Fairness in information access systems. Foundations and Trends® in Information Retrieval, 16(1-2), 1-177.

OUTLINE

- RECOMMENDER SYSTEMS (RS)
  - GROUP RECOMMENDER SYSTEMS (GRS)
- Explainable AI (XAI)
  - RS and GRS
- **Fairness**
  - **RS and GRS**
- IMPROVING Explainability and Fairness
  - Group Recommendation Systems
- CONCLUSIONS

FAIRNESS IN RECOMMENDER SYTEMS

- Recommender Systems
  - Successful tool for Enhancing Recommendations
  - Can have and strong influence
    - Information seen online
      - Social Media
    - Impact our beliefs, decision and acts
  - Can create business value for different stakeholders
    - Content Personalization and Engagement

FAIRNESS IN RECOMMENDER SYTEMS

- Recommender Systems
  - Fail to take into account critical aspects of recommendation
    - Fairness in one- and two-sided marketplaces
    - Biased behavior of algorithms
      - Towards certain groups of items and users
  - **Necessity of proposing fair and unbiased algorithms**
    - Ethical and Societal Implications
    - Regulatory and Transparency Concerns

FAIRNESS IN RECOMMENDER SYTEMS

- Raise Ethical questions: ***Fairness***

- Fairness in RS gain importance and increase attention
  - Historically RS benevolent systems create value for consumers
    - Support for finding useful items
  - Recently, awareness raised about negative effects
    - Promote items that gives more benefit to the platform or providers
    - Spread misinformation in social media
    - Echo Chambers and Filter Bubbles (Polarization, segmentation)
    - Etc.

FAIRNESS IN RECOMMENDER SYTEMS

NOTIONS OF FAIRNESS

- **Process Fairness**
  - The recommendations should be fair in process, which is also called procedural justice. If and only if the use of every one of the features in the set of features are fair
    - Ex: job recommendation, process fairness concerns whether the recommendation model is fair, concerns whether the recommendation model is fair, such as whether some unfair features (e.g., race) are used and whether the learned representations are fair.
- **Outcome Fairness**
  - The recommendations should lead to fair outcomes (distributive justice)
    - Ex: job recommendation, outcome fairness concerns the recommendation outcome, such as whether man would be more likely to be recommended than women even if they have the same ability.
  - The majority of existing research in recommendation focuses on the outcome fairness

FAIRNESS IN RECOMMENDER SYTEMS

- Outcome Fairness by target: **Group vs Individual**
  - ***One Sided Market***
    - *Minimizing the disparity between different user groups*
    - *Removing the algorithm's bias against the "protected" user group*
  - **Individual fairness** roughly expresses that similar individuals should be treated similarly, e.g., candidates with similar qualifications should be ranked similarly in a job recommendation scenario.

FAIRNESS IN RECOMMENDER SYTEMS

- Outcome Fairness by target: **Group vs Individual**

  - **Group fairness**, aims to ensure that “different groups have similar experience”, i.e., protected groups receive similar benefits from the decision-making as others. Typical groups are a majority or dominant group and a protected group (e.g., an ethnic minority)

    - The goal is to achieve some sort of *statistical parity* between *protected* groups.
    - The protected groups determined by characteristics as age, gender, or ethnicity.
    - Group fairness entails comparing, on average, the members of the privileged group against the unprivileged group

FAIRNESS IN RECOMMENDER SYTEMS

- **Group Fairness:** Aspects to consider
  - ***Two-Sided Market***
    - To protect not only the protected users, but also some item groups
    - Fairer approach also towards certain content providers
  - ***Benefit type*** (*exposure vs. relevance*)
    - Exposure relates to the degree to which items or item groups are exposed uniformly to all users/user groups.
    - Relevance (accuracy) indicates how well an item's exposure is effective, i.e., how well it meets the user's preference.

FAIRNESS IN RECOMMENDER SYTEMS

- **Group Fairness:** Aspects to consider
  - ***Major stakeholders*** (*consumer vs. providers*)
    - *Fairness evaluation* : users or items are splitted into non-overlapping groups (segments) based on attributes.
    - These attributes can be either supplied
      - Externally by the data provider (e.g., gender, age, race)
      - Or computed internally from the interaction data (e.g., based on user activity level, mainstreamness, or item popularity)

FAIRNESS IN RECOMMENDER SYTEMS

- **Group Fairness:** Aspects to consider
  - ***Major stakeholders*** (*consumer vs. providers*)
    - Most used attributes in the recommendation fairness, which can be utilized to operationalize the group fairness concept, are:
      - **Consumer fairness (C-Fairness)**: the disparate impact of recommendations on protected classes of consumers, associated with sensitive features, e.g., gender, race, and age.
      - **Producer Fairness (P-Fairness)**: ensure marketing diversity and avoid monopoly domination
      - **Combinations (CP-Fairness)** or multi-sided fairness

FAIRNESS IN RECOMMENDER SYTEMS

- Outcome Fairness by Concept
  - These concepts reflect researchers' understanding of what requirements should be met for fair outcomes.

Fairness Definitions		Abbr.	Description
Process Fairness		PF	the allocation process should be fair
Outcome Fairness		OF	the allocation outcome should be fair
Grouped by Target	Individual Fairness	IF	fairness should be guaranteed at the individual level
	Group Fairness	GF	fairness should be guaranteed at the group level
Grouped by Concept	Consistent Fairness	CO	similar individuals / different groups should receive similar outcomes
	Calibrated Fairness	CA	outcomes should be proportional to merits
	Counterfactual Fairness	CF	individuals should have the same allocation outcome in the real world as they do in the counterfactual world
	Rawlsian Maximin Fairness	RMF	the outcomes of the worst should be maximized
	Envy-free Fairness	EF	individuals should be free of envy
	Maximin-Shared Fairness	MSF	individuals / groups should get better outcomes than their maximin share

Research problems (overall)

- **Auditing RS fairness…**
- **Reducing RS algorithms unfairness….**

Our current focus at individual level fairness

- **C-fairness**

- **Assuring fairness across demographic features (e.g. the system/method performs similarly for any demographic class)**

Yao, S., & Huang, B. (2017, December). Beyond parity: fairness objectives for collaborative filtering. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 2925-2934).

Characterizing fairness

Unfairness metrics:

- **They measure the inconsistency in the estimation error, across both advantaged and disadvantaged users**

Unfairness value

$$U_{\text{val}} = \frac{1}{n}\sum_{j=1}^{n}\left|\left(\mathrm{E}_g\left[y\right]_j - \mathrm{E}_g\left[r\right]_j\right) - \left(\mathrm{E}_{\neg g}\left[y\right]_j - \mathrm{E}_{\neg g}\left[r\right]_j\right)\right| ,$$

Absolute unfairness

$$U_{\text{abs}} = \frac{1}{n}\sum_{j=1}^{n}\left|\left|\mathrm{E}_g\left[y\right]_j - \mathrm{E}_g\left[r\right]_j\right| - \left|\mathrm{E}_{\neg g}\left[y\right]_j - \mathrm{E}_{\neg g}\left[r\right]_j\right|\right| .$$

Yao, S., & Huang, B. (2017, December). Beyond parity: fairness objectives for collaborative filtering. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 2925-2934).

Characterizing fairness

Unfairness metrics:

Underestimation unfairness

$$U_{\text{under}} = \frac{1}{n}\sum_{j=1}^{n}\left|\max\{0, \mathrm{E}_g\left[r\right]_j - \mathrm{E}_g\left[y\right]_j\} - \max\{0, \mathrm{E}_{\neg g}\left[r\right]_j - \mathrm{E}_{\neg g}\left[y\right]_j\}\right| .$$

Overestimation unfairness

$$U_{\text{over}} = \frac{1}{n}\sum_{j=1}^{n}\left|\max\{0, \mathrm{E}_g\left[y\right]_j - \mathrm{E}_g\left[r\right]_j\} - \max\{0, \mathrm{E}_{\neg g}\left[y\right]_j - \mathrm{E}_{\neg g}\left[r\right]_j\}\right| .$$

Non-parity unfairness

$$U_{\text{par}} = \left|\mathrm{E}_g\left[y\right] - \mathrm{E}_{\neg g}\left[y\right]\right| .$$

Yao, S., & Huang, B. (2017, December). Beyond parity: fairness objectives for collaborative filtering. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 2925-2934).

Characterizing fairness

Unfairness metrics:

- **Beyond auditing…**
  - **Optimizing the fairness associated to the delivered recommendations…**

$$\min_{\boldsymbol{P},\boldsymbol{Q},\boldsymbol{u},\boldsymbol{v}} \; J(\boldsymbol{P},\boldsymbol{Q},\boldsymbol{u},\boldsymbol{v}) + U \; .$$

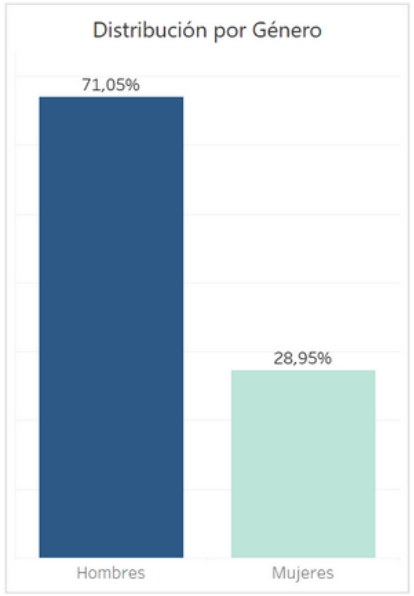
Yao, S., & Huang, B. (2017, December). Beyond parity: fairness objectives for collaborative filtering. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 2925-2934).

Case study

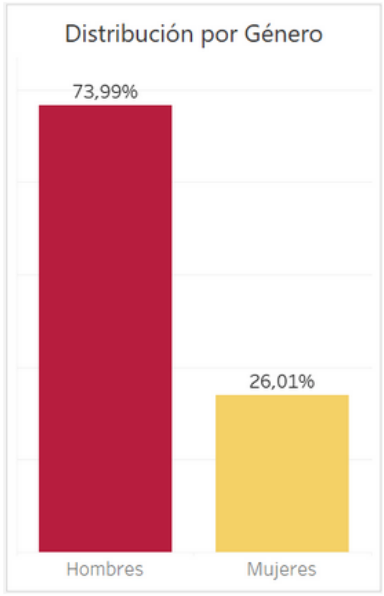
- **Task: Characterizing unfairness of two traditional RS approaches: UserKNN and MF-based.**
- **Datasets: Movielens100K (movies), and LastFM (music).**
- **Demographic features: Age and Gender**

Variable Género - Distribución

Movilens-100k



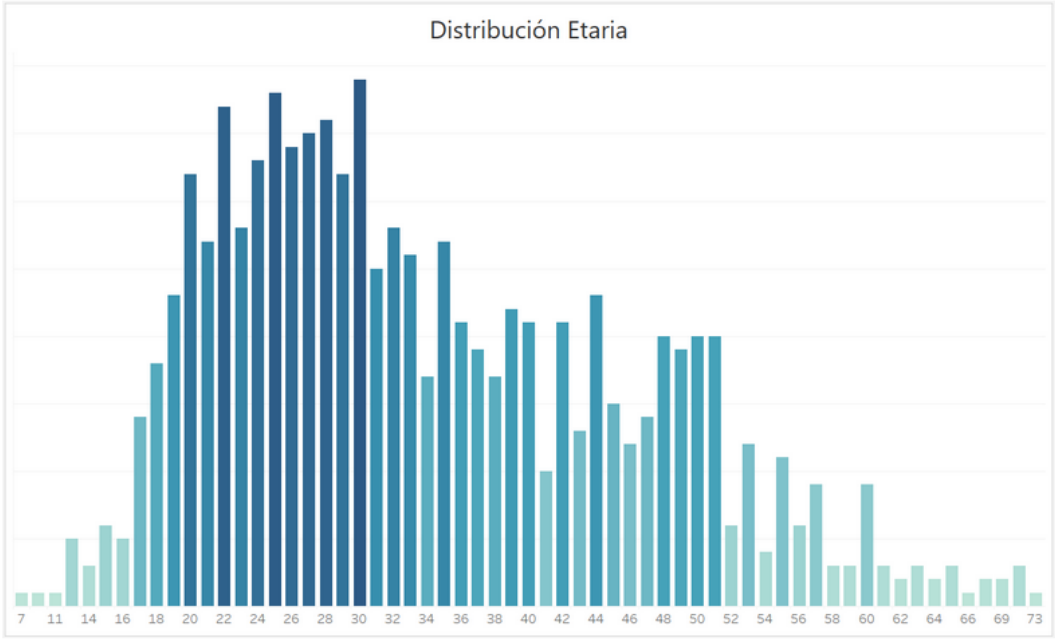
Last.fm-360K



“Escenario 1”
Advantaged users: “Hombres”
Disadvantaged users: “Mujeres”

“Escenario 2”
Advantaged users: “Mujeres”
Disadvantaged users: “Hombres”

Variable Edad - Distribución (Movielens - 100k)

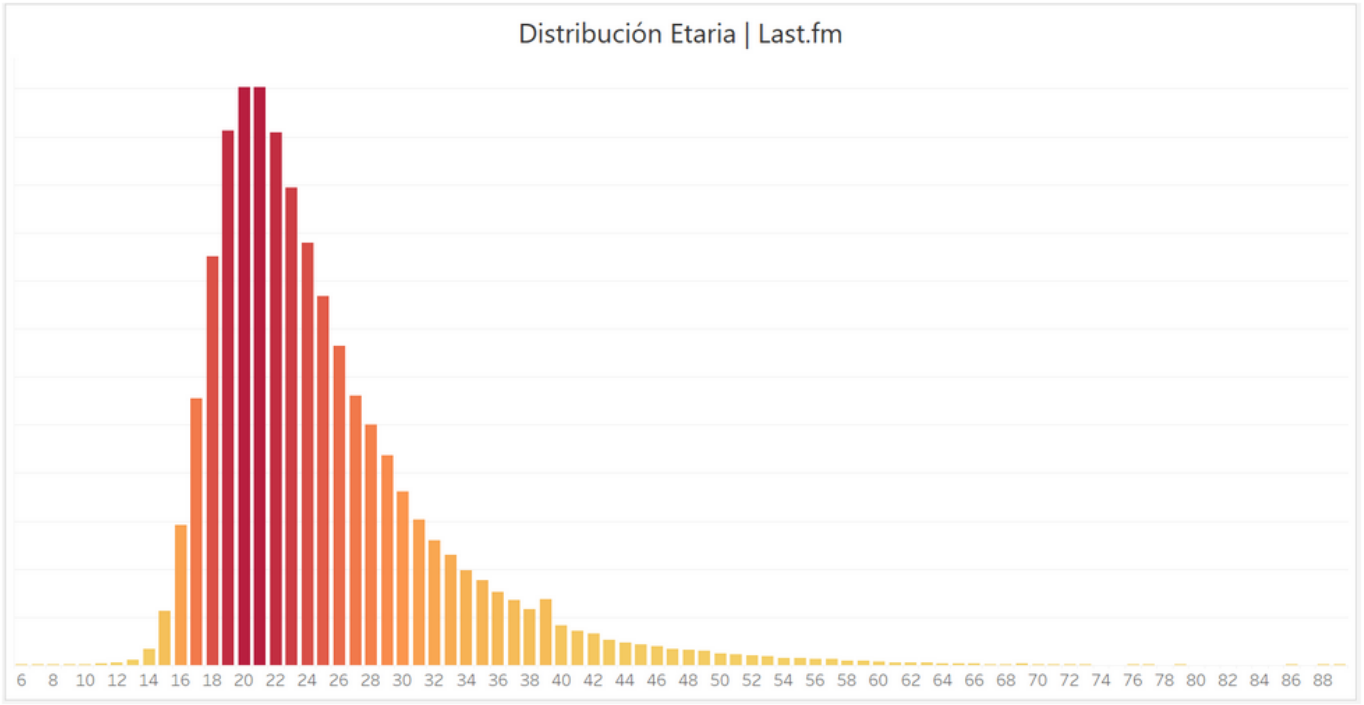


“Escenario 1”
Advantaged users: <=35
Disadvantaged users: >35

“Escenario 2”
Advantaged users: <=20
Disadvantaged users: >20

“Escenario 3”
Advantaged users: >=65
Disadvantaged users: <65

Variable Edad - Distribución (Last.fm - 360K)



“Escenario 1”
Advantaged users: <=25
Disadvantaged users: >25

“Escenario 2”
Advantaged users: <=17
Disadvantaged users: >17

“Escenario 3”
Advantaged users: >=65
Disadvantaged users: <65

Results

Dataset: Movielens-100k

Variable	Escenario	Algoritmo	Value Unfairness	Absolute Unfairness	Underestimation Unfairness	Overestimation Unfairness	Non Parity Unfairness
Edad	Escenario 1	KNN	0,4410	0,2832	0,2057	0,2353	0,1287
		MF	0,3989	0,2514	0,1853	0,2137	0,1124
	Escenario 2	KNN	0,3995	0,2721	0,2231	0,1764	0,1554
		MF	0,3690	0,2568	0,1716	0,1975	0,0124
	Escenario 3	KNN	0,0657	0,0523	0,0308	0,0349	0,0641
		MF	0,0652	0,0502	0,0308	0,0343	0,0771
Genero	Escenario 1	KNN	0,4724	0,3165	0,2228	0,2496	0,0101
		MF	0,4204	0,2811	0,1943	0,2262	0,0080
	Escenario 2	KNN	0,4724	0,3165	0,2228	0,2496	0,0101
		MF	0,4204	0,2811	0,1943	0,2262	0,0080

Overall, the unfair behavior tends to be more relevant for the gender variable; and for the first scenario of the age variable in the case of movies

Results

Dataset: Last.fm

Variable	Escenario	Algoritmo	Value Unfairness	Absolute Unfairness	Underestimation Unfairness	Overestimation Unfairness	Non Parity Unfairness
Edad	Escenario 1	KNN	0,1757	0,1199	0,0809	0,0948	0,0197
		MF	0,1475	0,1045	0,0733	0,0742	0,0170
	Escenario 2	KNN	0,1792	0,1267	0,0812	0,0980	0,0016
		MF	0,1496	0,1086	0,0755	0,0740	0,0092
	Escenario 3	KNN	0,0491	0,0416	0,0226	0,0264	0,0029
		MF	0,0409	0,0357	0,0212	0,0197	0,0162
Genero	Escenario 1	KNN	0,1679	0,1186	0,0742	0,0937	0,0022
		MF	0,1429	0,1025	0,0715	0,0713	0,0181
	Escenario 2	KNN	0,1679	0,1186	0,0742	0,0937	0,0022
		MF	0,1429	0,1025	0,0715	0,0713	0,0181

In the case of the music, the second scenario for “Age”, also has a higher unfairness value

Fairness in GRS.

- Starting point

  According to Deldjoo et al (2023):

  Most other works that focus on individual fairness address problems of *group recommendation*, i.e., situations where a recommender is used to make item suggestions for a group of users…..

  There is a lack of research focused on group fairness considering protected classes in group recommendation

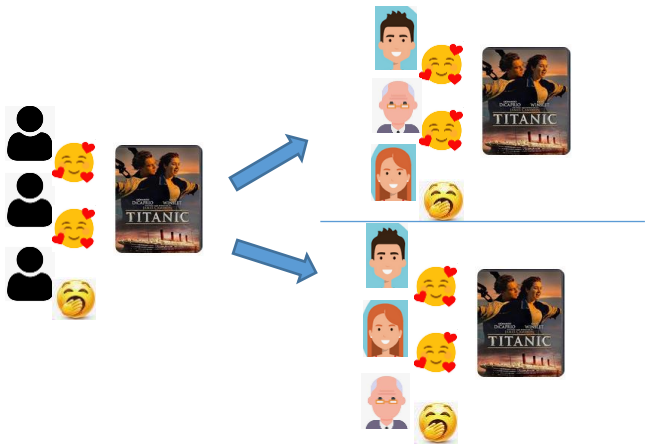
Y. Deldjoo, D. Jannach, A. Bellogin, A. Difonzo and D. Zanzonelli. Fairness in recommender systems: research landscape and future directions, User Modeling and User-Adapted Interaction , 24, 59-108, 2024.

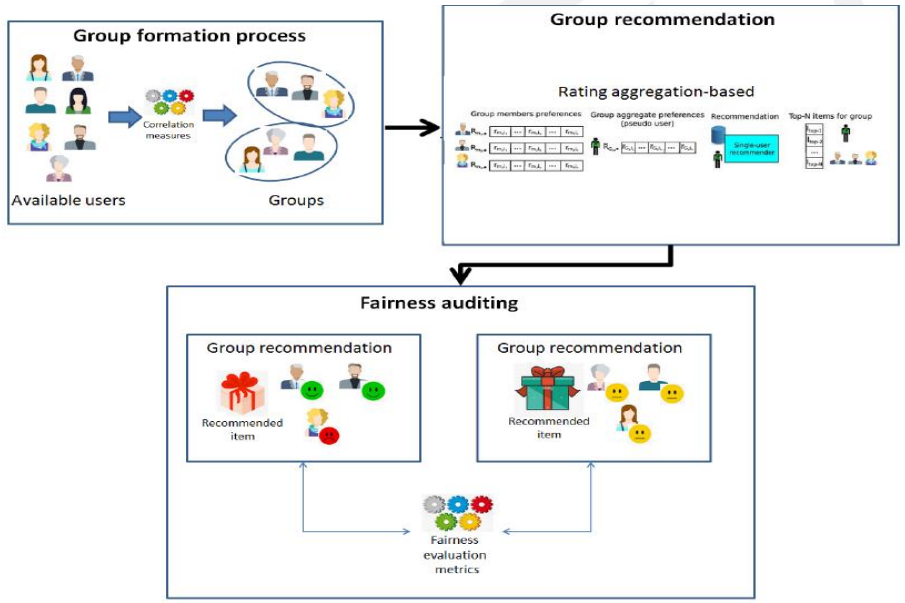
Fairness in GRS.

- Multi stakeholder scenario
  - *Consumer fairness (C-Fairness)*
  - Producer Fairness (P-Fairness)
  - Combinations (CP-Fairness)
- Protected classes of sub-groups
- Current fairness GRS proposals
  - Only by optimizing the exposure of items in the recommendation list
    - Position of the item
  - Without user's demographic attributes
    - Key information for characterizing algorithmic fairness

Fairness in GRS.

- How to improve Fairness in GRS
  - Introducing demographic attributes-related fairness





Yera, R., Barranco, M. & Martínez L. (2024). A Novel Approach for Measuring Demographic Parity Fairness in Group Recommendation. In *Intelligent Management of Data and Information in Decision Making* (pp. 195—202). World Scientific Co.

Fairness in GRS.

- Measuring Fairness in GRS

1. Identify deviation of the actual user preferences r_{ui}, wrt the predictions r_{Gi} done for the group using a GRS:

$$User\_Group\_Dev(u) = \frac{1}{|R_u|} \sum_{r_{ui} \in R_u} (r_{ui} - r_{Gi})$$

2. Characterize the recommendation unfairness *Unfairn* of user u in a GRS framework, as how close is his/her User_Group_Dev(u) value, wrt User_Group_Dev(v) values of the other users v ∈ G

$$Unfairn(u) = min(|User\_Group\_Dev(u) - User\_Group\_Dev(v)|); v \in G$$

3. The unfairness level associated to a group G is represented as the absolute difference between the average unfairness of the users respectively belonging to the set A and D of the advantaged and disadvantaged users in the group G.

$$Unfairn(G) = |\frac{1}{|A|} \sum_{u \in A} Unfairn(u) - \frac{1}{|D|} \sum_{v \in D} Unfair(v)|$$

Fairness in GRS

- Experimental Settings
  - Disadvantaged demographic user classes:
    - Gender=Woman
    - Age<25
    - Age>55
  - Movielens 100K dataset and Rating aggregation-based GRS
  - Groups of 4 members, 3 in the advantaged and 1 in the disadvantaged class.
  - The pairwise Pearson correlation similarity between all the group members, satisfies $\alpha > 0$.

  - The goal of the study is:
    - To characterize the Unfair values for each group member.
    - To characterize the Unfair values for the whole group.

Fairness in GRS. Findings

Table 1. Average unfairness values according to Equation 2, for disadvantaged and advantaged classes in group members

	Gender	Age	Age
Disadvantaged users	$Gender = Woman$	$Age < 25$	$Age > 55$
	0.2454±0.1893	0.1572±0.1202	0.1866±0.1476
Advantaged users	$Gender = Man$	$Age >= 25$	$Age <= 55$
	0.2070±0.1469	0.2007±0.1516	0.2312±0.1573

Table 2. Average unfairness values for the identified groups according to Equation 3

	Gender	Age	Age
Group unfairness	$Gender = Woman$	$Age < 25$	$Age > 55$
	0.1790±0.1029	0.1455±0.0741	0.2037±0.1291

- Table 1: Unfairness values for individuals in each class.
  - Larger unfairness values linked to the gender attribute.
  - In the case of the age criteria, it was obtained a higher unfairness value for the advantaged class in relation to the disadvantaged class.

Fairness in GRS. Findings

Table 1. Average unfairness values according to Equation 2, for disadvantaged and advantaged classes in group members

	Gender	Age	Age
Disadvantaged users	$Gender = Woman$	$Age < 25$	$Age > 55$
	0.2454±0.1893	0.1572±0.1202	0.1866±0.1476
Advantaged users	$Gender = Man$	$Age >= 25$	$Age <= 55$
	0.2070±0.1469	0.2007±0.1516	0.2312±0.1573

Table 2. Average unfairness values for the identified groups according to Equation 3

	Gender	Age	Age
Group unfairness	$Gender = Woman$	$Age < 25$	$Age > 55$
	0.1790±0.1029	0.1455±0.0741	0.2037±0.1291

- Table 2: Unfairness values at the group level.
  - The larger group unfairness was associated to Age > 55.
  - In the case of the gender, here there is a lower unfairness taking into account that in Table 1, users from both classes have closer Unfairn values.

CONCLUSIONS

- EXPLAINABILITY
  - Ethical concept required in AI based systems, fairly important in RS
    - Explainability plays a crucial role in building user trust by helping users understand why certain items are recommended
    - Explanations enable users to make more informed and confident decisions
    - Explanations can increase user engagement by persuading them to explore or purchase recommended items
    - Explainability provides transparency into model behavior, facilitating debugging, auditing, and compliance

CONCLUSIONS

- FAIRNESS
  - In RS is a **multidimensional challenge** that cannot be addressed with a one-size-fits-all solution

    - Different fairness metrics capture different types of bias, and optimizing for one may worsen others.
    - There are trade-offs between fairness and accuracy
    - User group characteristics (e.g., size,) significantly affect fairness outcomes.
    - No single algorithm achieves fairness across all metrics and contexts, underscoring the need for context-aware fairness strategies.

THANKS A LOT
FOR YOUR ATTENTION

QUESTIONS