

# How Dataset Characteristics Affect the Robustness of Collaborative Recommendation Models

*Deldjoo Yashar, Di Noia Tommaso, Di Sciascio Eugenio, Merra Felice Antonio*

*{yashar.deldjoo,tommaso.dinoia,eugenio.disciascio,felice.merra}@poliba.it*

The 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval  
July 25-30, 2020 (Xi'an, China)



# Outline

1. Introduction and Preliminaries
2. Problem Formalization
3. Experimental Settings
4. Results and Discussion
5. Conclusion and Future Works

# 1. Introduction and Preliminaries

Recommender Systems (RS) support users' decision-making process in online or e-commerce platforms (e.g., Netflix, Zelando and Amazon.com).



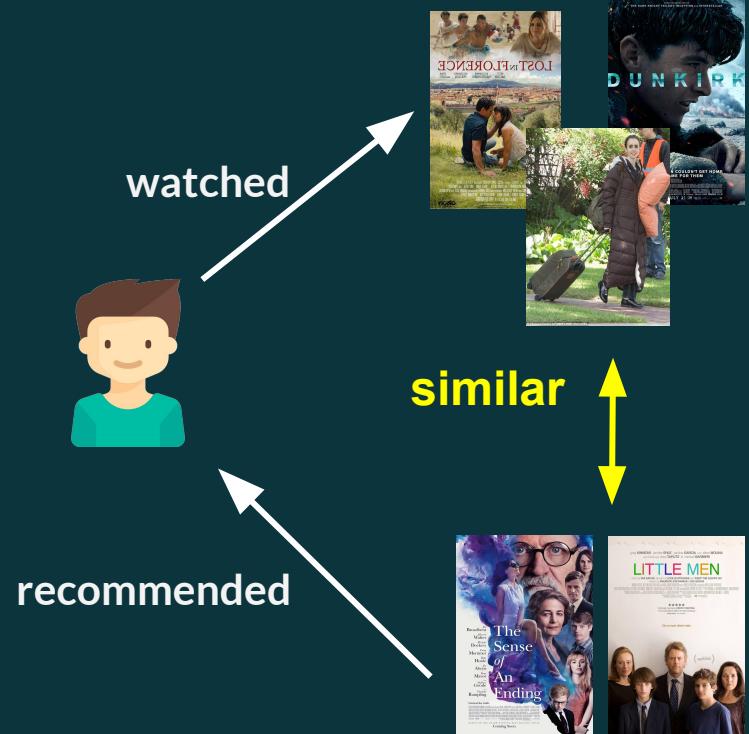
# Main Classes of Recommendation

- **Content-based filtering (CBF)**

Recommend products based on **similarity** between user profile and unseen items

## Main content-based similarity types

- Editorial metadata: genre, artists
- User generated: tags, reviews
- Semantic data: wikidata, DBpedia [1]
- Multimedia: audio, visual content [2]



[1] Oramas et al., "Sound and music recommendation with knowledge graphs." ACM TIST (2020)

[2] Deldjoo et al., "Recommender Systems Leveraging Multimedia Content." ACM Computing Surveys (CSUR) (2020)

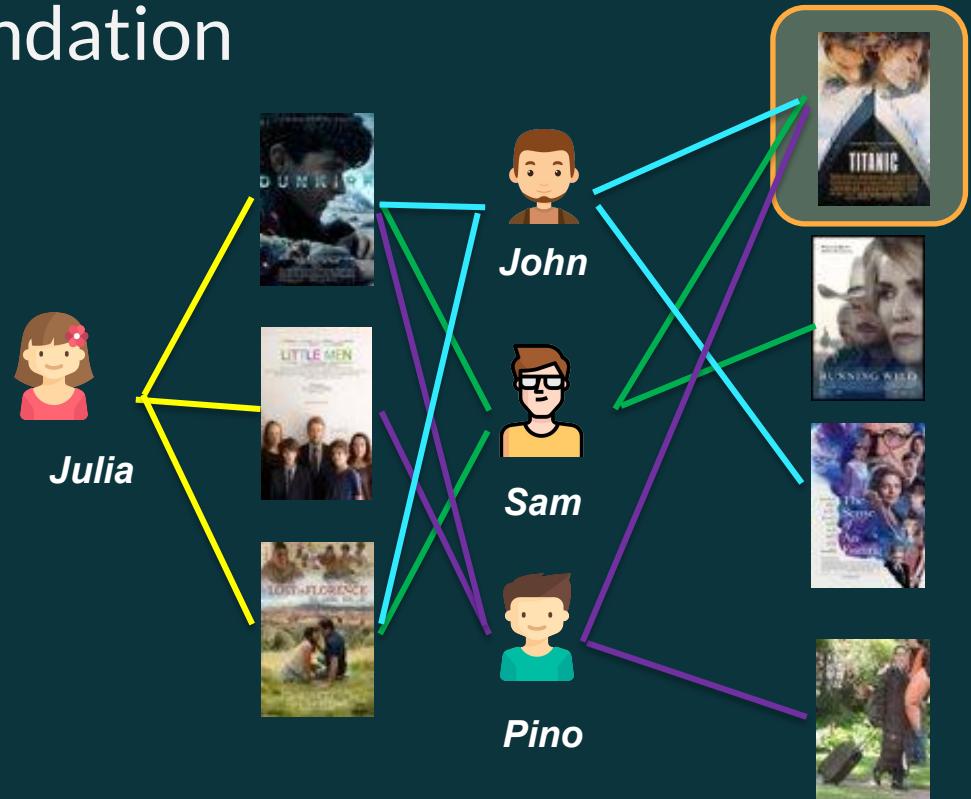
# Main Classes of Recommendation

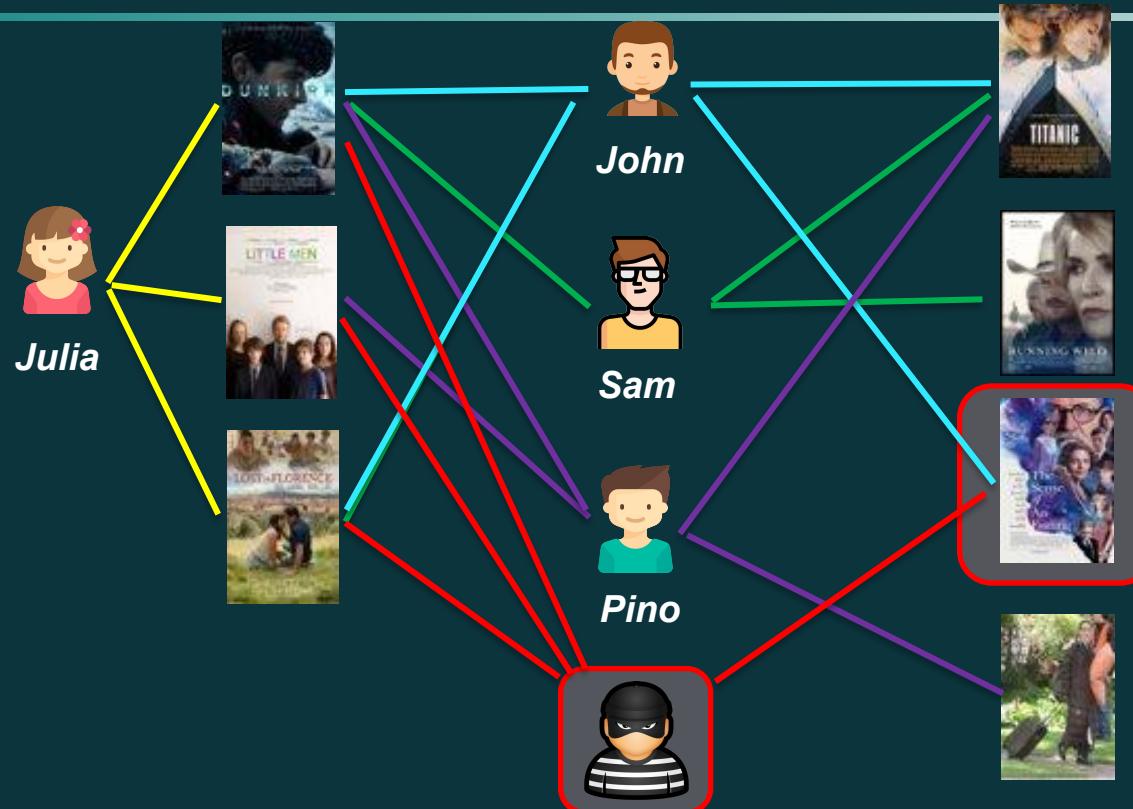
- **Collaborative filtering (CF)**

Suggest products experienced by similar users.

Main types of CF models

- Model-based: MF, FM, DCN
- Memory-based: item-knn, user-knn





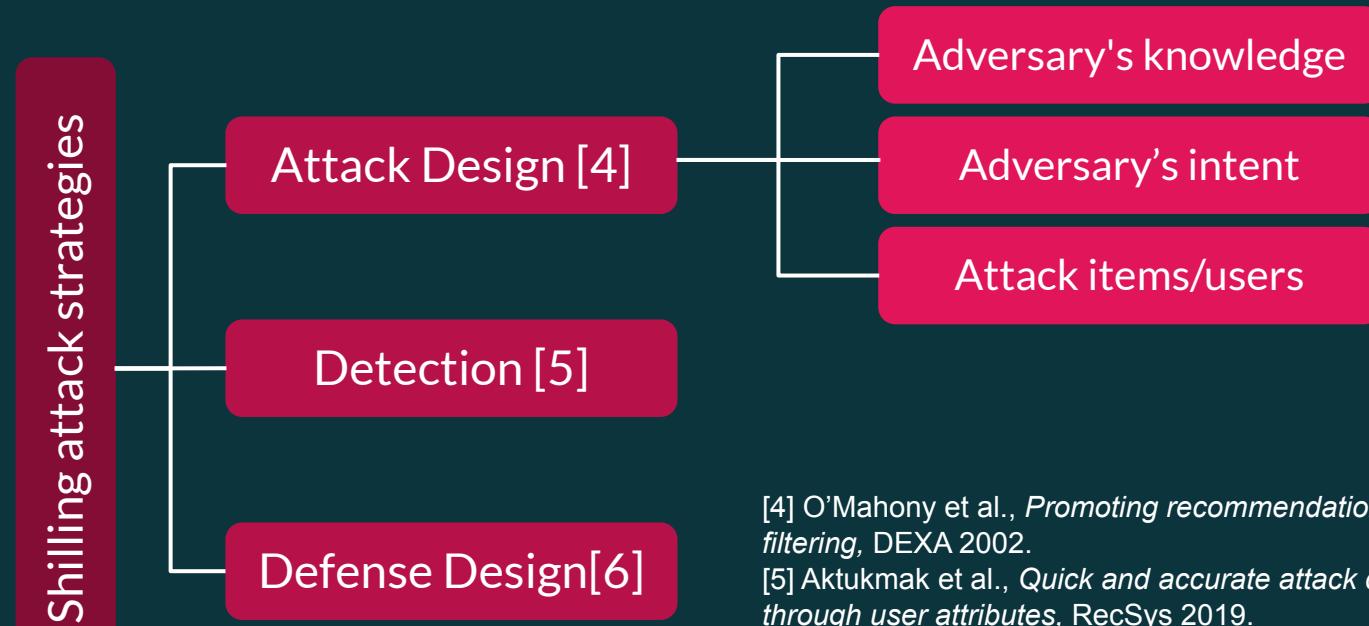
CF models are  
vulnerable against  
manually crafted  
**SHILLING  
PROFILES**

[3] Gunes et al., *Shilling attacks against recommender systems: a comprehensive survey*, Artif. Intell. Rev. 42, 4 (2014)

# Goals of Malicious Attacks

- Business
  - Personal gain against a competitor
  - Market penetration
- Politics
  - Fake social media accounts to spread news about a specific party or belief system
- Privacy
  - Attack privacy of users, data leakage
- Others
  - Attack fairness of a recommendation system
  - Reduce trustworthiness of the online platform

# Prior researches in shilling attack

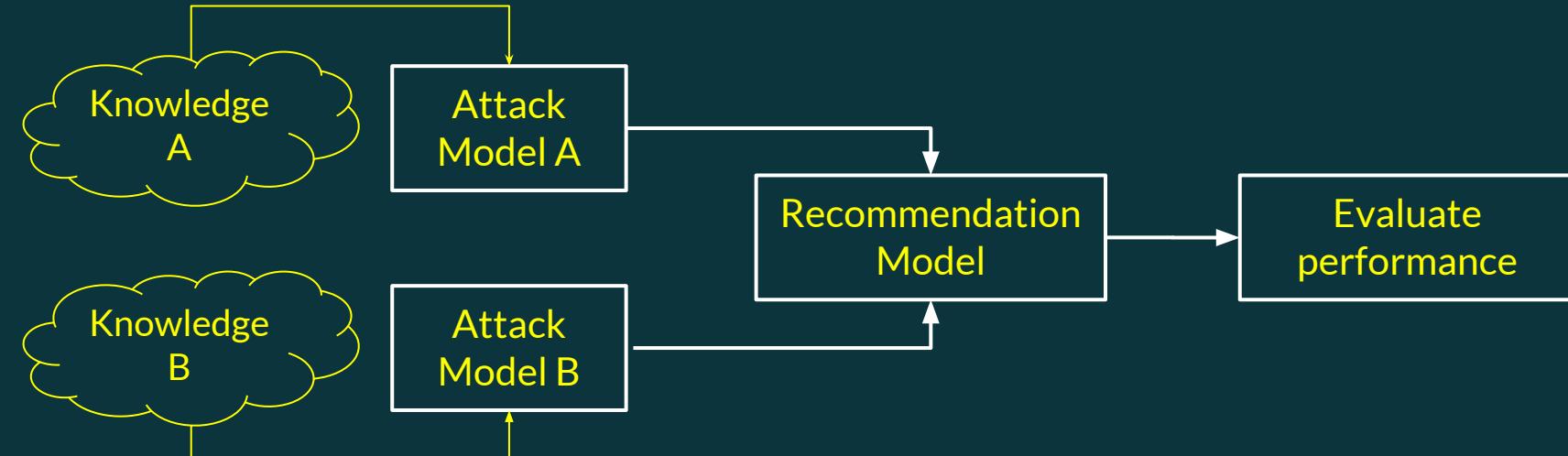


[4] O'Mahony et al., *Promoting recommendations: An attack on collaborative filtering*, DEXA 2002.

[5] Aktukmak et al., *Quick and accurate attack detection in recommender systems through user attributes*, RecSys 2019.

[6] Zhang et al., *Robust collaborative filtering based on non-negative matrix factorization and R1-norm*, Knowl.-Based Syst 2017

# Previous Studies

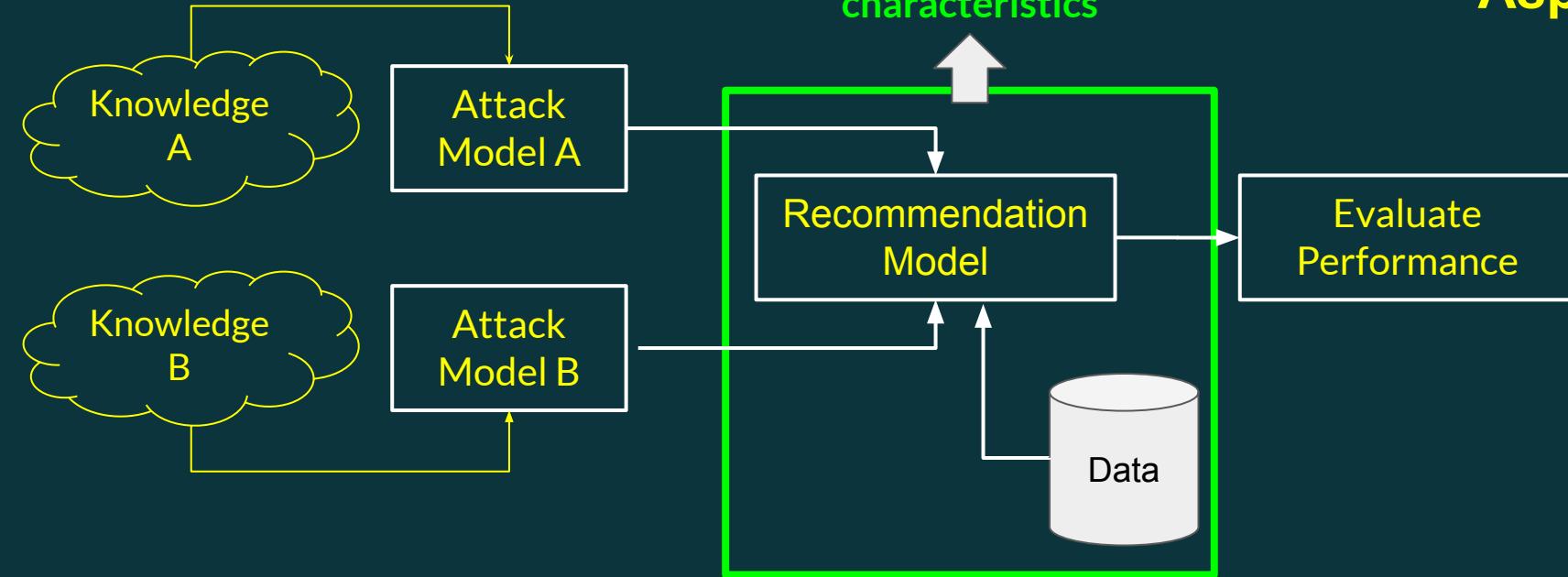


1. Which **attack models** impact more the performance of certain recommendation models?
2. Which **amount of knowledge** on a rec. model is required for specific attack to influence a recommendation algorithm?

# Previous Studies

Solid understanding of  
the Impact of **dataset  
characteristics**

# Forgotten Aspect



# Main Research Question

*Given popular shilling attack types and CF models already recognized by the community, which **dataset characteristics** can explain an observed change in the performance of recommendation?*

# The Main Contributions

1. **Modeling.** We studied the influence of data characteristics on the recommendation performance using a **regression-based explanatory model** (inspired by [7])
2. **Data characteristics.** We validate the correlation between data characteristics and attack effectiveness on an **extensive suite of data characteristics**
3. **Experiments.** We conducted an empirical analysis on:
  - o 6 Shilling Attack Strategies
  - o 3 Collaborative Filtering models
  - o 3 Real-World datasets

[7] Adomavicius and Zhang, *Impact of data characteristics on recommender systems performance*, ACM TIST 2012.

## 2. Problem Formalization

# The Independent Variables (IVs)

- The IVs are the dataset characteristics under investigation.
- We investigated 6 IVs categorized as follows:
  - IVs based on URM structure (**Structural**)
  - IVs based on rating frequency of the URM (**Distributional**)
  - IVs based on rating values of the URM (**Value-based**)

# Structural IVS

- Space Size

$$x_1 = \log_{10} \left( \frac{|\mathcal{U}| \cdot |\mathcal{I}|}{sc} \right)$$

$|\mathcal{I}|$  = Num. of Items

- Shape

$$x_2 = \log_{10} \left( \frac{|\mathcal{U}|}{|\mathcal{I}|} \right)$$

Scaling factor

- Density

$$x_3 = \log_{10} \left( \frac{|\mathcal{K}|}{|\mathcal{U}| \times |\mathcal{I}|} \right)$$

$|\mathcal{U}|$  = Num. of Users

$|\mathcal{K}|$  = Num. of Ratings

Log transformation to normalize the distribution of the variables.

[8] Deldjoo et al., Assessing the Impact of a User-Item Collaborative Attack on Class of Users, In ImpactRS@RecSys' 19

# Distributional IVs

$|\mathcal{K}_i|$  = Num. of Ratings Received by Item  $i$

$|\mathcal{K}_u|$  = Num. of Ratings Given by User  $u$

- Gini Index for Item

$$x_4 = 1 - 2 \sum_{i=1}^{|\mathcal{I}|} \left( \frac{|\mathcal{I}|+1-i}{|\mathcal{I}|+1} \right) \times \left( \frac{|\mathcal{K}_i|}{|\mathcal{K}|} \right)$$

- Gini Index for Users

$$x_5 = 1 - 2 \sum_{u=1}^{|\mathcal{U}|} \left( \frac{|\mathcal{U}|+1-u}{|\mathcal{U}|+1} \right) \times \left( \frac{|\mathcal{K}_u|}{|\mathcal{K}|} \right)$$

Gini coefficients = 0 --> Equal Popularity (e.g., all users give the same number of ratings)

Gini coefficients = 1 --> Total Inequality (e.g., only one user has given all ratings)

[9] Herlocker et al., *Explaining collaborative filtering recommendations*, In CSCW 2000

# Value-based IVs

- Standard Deviation of Rating Values

$$x_6 = \sqrt{\frac{\sum_{i=1}^{|\mathcal{K}|} (r_i - \bar{r})^2}{|\mathcal{K}| - 1}}$$

where  $r_i$  is the i-th Rating, and  $\bar{r}$  is the Average Rating Value.

# The Dependent Variables (DV)

- The dependent variable (DV) represents the effectiveness of the attack on RS.
- Inspired by the Overall Hit Ratio[10], we proposed and investigated the Incremental Overall Hit Ratio:

Let  $HR@k(\mathcal{I}_T, \mathcal{U}_T) = \frac{\sum_{i_t \in \mathcal{I}_T} hit(i_t, \mathcal{U}_T)}{|\mathcal{I}_T|}$  then  $\Delta_{HR@k} = \hat{HR}@k - HR@k$

Before the Attack

Hit Function

After the Attack

[10] Charu C. Aggarwal, *Recommender Systems - The Textbook*, Springer 2016

# The Explanatory Framework (EF)

- The EF tests the *causal hypothesis* in a theoretical construct:  
*Are a set of effects measured by IVs the cause for an effect measured by the DV?*
- Our Causal Hypothesis:  
*Are the data characteristics causing variations in attack performance?*
- Inspired by Adomavicious et al.[7], we use a **regression model** as the interpretable model.

# The Regression Model (Compact Form)

- The regression model used to study the causal relationship is

$$\mathbf{y} = \epsilon + \theta_0 + \theta_d \mathbf{X}_d + \theta_c \mathbf{X}_c$$

where

$\theta_0$  represents the expected value of  $\mathbf{y}$

$\theta_d = [\theta_1, \theta_2, \dots, \theta_{D-1}]$  is the vector containing coefficients of the dummy variable  $\mathbf{X}_d$

$\theta_c = [\theta_1, \theta_2, \dots, \theta_C]$  is the vector of the regression coefficient associated with the IVs

$\mathbf{X}_c$  is the matrix containing the IVs values

# The Explanatory Analysis

- We applied the EF to for two analysis
  - **Within-dataset analysis.** Study <Dataset, Attack, CF-RS> combinations

$$(\theta_0^*, \boldsymbol{\theta}_c^*) = \min_{\theta_0, \boldsymbol{\theta}_c} \frac{1}{2} \|\mathbf{y} - \theta_0 - \boldsymbol{\theta}_c \mathbf{X}_c\|_2^2$$

- **Between-dataset analysis.** Study <Attack, CF-RS> combinations

$$(\theta_0^*, \boldsymbol{\theta}_d^*, \boldsymbol{\theta}_c^*) = \min_{\theta_0, \boldsymbol{\theta}_d, \boldsymbol{\theta}_c} \frac{1}{2} \|\mathbf{y} - \theta_0 - \boldsymbol{\theta}_d \mathbf{X}_d - \underline{\boldsymbol{\theta}_c \mathbf{X}_c}\|_2^2$$

dummy term for the  
dataset-independent analysis

# 3. Experimental Settings

# Datasets

Dataset	$ \mathcal{U} $	$ \mathcal{I} $	$ \mathcal{K} $	$density$
ML-20M	138,493	26,744	20,000,263	0.0054
Yelp	25,677	25,778	705,994	0.0010
LFM-1b	120,175	521,232	25,285,767	0.0004

# CF Recommender Models

- **User-kNN [11]**: predicts the score of unknown user-item pairs by considering the feedback of the users in the neighborhood.
- **Item-kNN [11]**: estimates the user-item rating score by using the recorded user's feedback on the neighborhood items.
- **Matrix Factorization /SVD [12]**: learns user-item preferences, by factorizing the sparse user-item feedback matrix.

[11] Koren, *Factor in the neighbors: Scalable and accurate collaborative filtering*, TKDD 2010

[12] Koren et al., *Matrix factorization techniques for recommender systems*, IEEE Computer 2009

# Shilling Attacks

Taxonomy based on [13]:

- **INTENT**
  - **PUSH** (Increase the probability of a target item to be recommended)
  - **NUKE** (Reduce the probability of a victim item to be recommended)
- **KNOWLEDGE**
  - **Low-Knowledge**: attackers require little or no knowledge about the rating distribution
  - **Informed**: adversaries get knowledge on dataset rating distribution

[13] Lam, S.K., Riedl, J., *Shilling recommender systems for fun and profit*, WWW 2004

# The Form of Fake Profiles

$I_S$	$I_F$			$I_\emptyset$			$I_T$		
$i_s^{(1)}$	$\dots$	$i_s^{(\alpha)}$	$i_f^{(1)}$	$\dots$	$i_f^{(\phi)}$	$i_\emptyset^{(1)}$	$\dots$	$i_\emptyset^{(\chi)}$	$i_t$

$I_S$  Items selected in case of informed strategies, which exploit attacker's knowledge.

$I_F$  Items **RANDOMLY** selected to make the *shilling profile* difficult to be detected.

$I_\emptyset$  Items that will not contain any ratings in the profile **Dependent on the Attack Strategy**

$I_T$  **Target Item** attacked to change. (Rating = 5 for *push intent*, 1 for *nuke intent*)

[14] Bhaumik et al., *Securing collaborative filtering against malicious attacks through anomaly detection*, ITWP 2016

# The Attack Strategies

Attack Type	$I_S$		$I_F$		$I_\phi$	$I_T$
	Items	Rating	Items	Ratings		
Random	$\emptyset$		$\frac{\sum_{u \in U}  I_u }{ U } - 1$	$rnd(N(\mu, \sigma^2))$	$I - I_F$	$max$
Love-Hate	$\emptyset$		$\frac{\sum_{u \in U}  I_u }{ U } - 1$	$min$	$I - I_F$	$max$
Bandwagon	$(\frac{\sum_{u \in U}  I_u }{ U })/2 - 1$	$max$	$(\frac{\sum_{u \in U}  I_u }{ U })/2$	$rnd(N(\mu, \sigma^2))$	$I - I_S - I_F$	$max$
Popular	$\frac{\sum_{u \in U}  I_u }{ U } - 1$	$min \text{ if } \mu_f < \mu \text{ else } min + 1$	$\emptyset$		$I - I_S$	$max$
Average	$\emptyset$		$\frac{\sum_{u \in U}  I_u }{ U } - 1$	$rnd(N(\mu_f, \sigma_f^2))$	$I - I_F$	$max$
P. Knowledge	$\frac{\sum_{u \in U}  I_u }{ U } - 1$	$max$	$\emptyset$		$I - I_S$	$max$

# Sub-Sample generation procedure

**Input:** URM

**Results:**  $\mathcal{N}$  sub-datasets ( $urm_n$ )

$n \leftarrow 1$

**while**  $n \leq \mathcal{N}$  **do**

    Random shuffle the row of the URM

$num_{users} \leftarrow rnd([100, 2500])$

$num_{items} \leftarrow rnd([100, 2500])$

$urm_n \leftarrow$  Selection of  $num_{users}$ ,  $num_{items}$  from URM

**if**  $density(urm_n) \in [0.0005, 0.01]$  **then**

$n \leftarrow n + 1$

# The Evaluation

To evaluate the EF we studied:

- **Adjusted Coefficient of Determination  $R^2$** 
  - 1 -> The DV is completely explained by the IVs
  - 0 -> The model explains none of the variability in the output
- **Directionality of the Regression Coefficients.**
  - +/- -> Positive/Negative Impact of the IV on the DV
- **Significance of the Regression Coefficients**
  - $p < 0.05$  Statistically Significant Results

# Evaluation Questions

1. Is there an **underlying relationship** between the IVs and the effectiveness of shilling attacks measured in terms of Overall Hit Ratio, the DV?
2. How **significant** is the impact of each IV? Is the **directionality** positive or negative?
3. Is the impact **consistent** in a domain-independent setting?

# 4. Results and Discussion

# Within Dataset Analysis: Coefficient of Determination

- Given a <Dataset, Attack, CF-model> we observed that the six IVs can explain more than 60% of the DV variation

$\Delta_{HR@10}$	User-kNN			
	ML-20M	Yelp	LFM-1b	
Random	$R^2(\text{adj.} R^2)$	0.761(0.758)	0.838(0.835)	0.673(0.668)
	<i>Constant</i>	.179***	.609***	.717***
	<i>SpaceSize<sub>log</sub></i>	-0.063***	.041	-0.629***
	<i>Shape<sub>log</sub></i>	.184***	.248***	.288*
	<i>Density<sub>log</sub></i>	-0.189***	-0.316*	-1.546***
	<i>Gini<sub>users</sub></i>	.277	-0.012	1.901***
	<i>Gini<sub>item</sub></i>	-0.102	-0.485	1.753***
	<i>Std<sub>rating</sub></i>	-0.072	.287	-0.152



- Maximum** values for the SVD model on Yelp (>85%)
- Minimum** on User-kNN for LFM-1b (from 66% to 67%).

# Within Dataset Analysis: Significance

- The significance of the regression coefficients varies for group of IVs.
- The coefficients computed for the **Structural Characteristics** are **mostly significant**.
- Gini indices** coeff. are **mostly significant** for shilling attacks against **SVD** (Yelp, LFM)
- Standard Deviation** coeff. are generally **NOT Significant** ( $p\text{-value}>0.05$ )

$\Delta_{HR@10}$	SVD			
	ML-20M	Yelp	LFM-1b	
Bandwagon	$R^2(\text{adj.}R^2)$	0.841(0.839)	0.914(0.912)	0.786(0.784)
	<i>Constant</i>	.435***	.522***	.689***
	<i>SpaceSize<sub>log</sub></i>	-0.006	.372***	-0.366***
	<i>Shape<sub>log</sub></i>	.244***	.278***	.206*
	<i>Density<sub>log</sub></i>	-0.314***	.401***	-1.047***
	<i>Gini<sub>users</sub></i>	.602***	-0.680**	.976*
	<i>Gini<sub>item</sub></i>	.268	-1.278***	1.276***
	<i>Std<sub>rating</sub></i>	-0.290	.321*	-0.066

\*\*\* $p \leq .001$ , \*\* $p \leq .01$ , \* $p \leq .05$

# Within Dataset Analysis: Directionality

- **Density** and **Space** have **Negative Impact**.

For instance, **Increasing** the **density** (or decreasing sparsity) of the dataset **REDUCES** the attacks' effectiveness.

- **Shape** has **Positive Impact**:

Increasing the shape leads to have more users than items. Pushing the target item might be simpler since there are few items to overcome considering a fixed size and density.

# Between Dataset Analysis

To provide a **domain-independent analysis** by combining all the sub-samples of the 3 datasets and check the **CONSISTENCY** of the previous results.

		$\Delta_{HR@10}$	User- <i>k</i> NN	Item- <i>k</i> NN	SVD
Average	$R^2(\text{adj.}R^2)$	0.828(0.827)	0.810(0.809)	0.844(0.843)	
	ML-20M (Constant)	.187***	.275***	.502***	
	Yelp	.421***	.332***	.020***	
	LFM-1b	.529***	.438***	.186***	
	$SpaceSize_{log}$	-0.193***	-0.082***	.065***	
	$Shape_{log}$	.152***	.107***	.192***	
	$Density_{log}$	-0.718***	-0.522***	-0.219***	
	$Gini_{user}$	.559***	-0.039	.011	
	$Gini_{item}$	.717***	.407***	-0.062	
	$Std_{rating}$	-0.054	.059	-0.013	

\*\*\* $p \leq .001$ , \*\* $p \leq .01$ , \* $p \leq .05$

# Between Dataset Analysis: Discussion

- The **coefficients of determination** are **consistent** with those in within-dataset analysis in most experimental cases
- Results still support that **structural URM properties** have a **statistically significant impact** on each CF model ( $p$ -values  $< 0.001$ )
- The **directionality** analysis of structural IVs is **consistent** with the insights drawn from the within dataset analysis.

# 5. Conclusion and Future Works

# Conclusion

- We studied the impact of data characteristics on the effectiveness of most famous shilling attacks against popular CF methods with a regression model.
- The structural, distributional, and value-based properties:
  - Account for the variations in attack performance (**global perspective**)
  - Have differences in the significance, and directionality (**local perspective**).
- **We plan to extend:**
  - The set of studied characteristics (e.g., user-item relations)
  - CF models (e.g., **deep learning** approaches)
  - **Novel attacks such as adversarial machine-learned attacks**

## Contact Presenter Info:



[felice.merra@poliba.i  
t](mailto:felice.merra@poliba.it)



[@merrafelice](https://twitter.com/merrafelice)

