My friends also prefer diverse music: homophily and link prediction with user preferences for mainstream, novelty, and diversity in music

Tomislav Duricic, Dominik Kowald, Markus Schedl, and Elisabeth Lex

Content

1. **Motivation and research objectives**
2. **Data and preprocessing**
3. **Homophily in user preferences**
4. **Link prediction with user preferences**
5. **Summary and conclusion**
6. **Contact Page**

Motivation & Background

Homophily

*„**<u>Similarity breeds connection</u>**." **Homophily*** [1] *is a term that describes this tendency of people to connect with others that are similar to themselves in some aspect. Such aspects include **sociodemographic**, **behavioral**, or **intrapersonal characteristics*** [2].

[1] McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. Annual review of sociology, 27(1), 415-444.
[2] Block, P., & Grund, T. (2014). Multidimensional homophily in friendship networks. Network Science, 2(2), 189-212.

Music plays an important role in friendship formation

- Tarrant et al. [3] have found in-group favoritism in male adolescent peer groups along musical preference dimensions and that music preferences in friendship groups may affect adolescent music preferences
- Selfout et al. [4] have shown that non-mainstream music similarity and similarity in the overall patterning of music preferences both are related to friendship formation, but not to friendship stability
- Zillmann et al. [5] have even shown that similarity in music taste is associated with closeness and relationship satisfaction

[3] Tarrant, M., North, A. C., & Hargreaves, D. J. (2001). Social categorization, self-esteem, and the estimated musical preferences of male adolescents. The Journal of social psychology, 141(5), 565-581.
[4] Selfhout, M. H., Branje, S. J., ter Bogt, T. F., & Meeus, W. H. (2009). The role of music preferences in early adolescents' friendship formation and stability. Journal of adolescence, 32(1), 95-107.
[5] Zillmann, D., & Bhatia, A. (1989). Effects of associating with musical genres on heterosexual attraction. Communication Research, 16(2), 263-288.

What defines one’s music taste?

- „*Musical taste is often paired together with musical preferences. Most people understand musical taste as one’s preferences in music: particular genres, styles, artists”* [6]
- However, music listeners can also be described by their preferences towards **mainstream**, **diverse**, or **novel** music content [7]
- Using features capturing those preferences for user modeling has proven effective in a music recommendation setting

[6] http://serious-science.org/musical-taste-9601
[7] M. Schedl and D. Hauger (2015), Tailoring music recommendations to users by considering diversity, mainstreaminess, and novelty. SIGIR.

Existing research

- Previous studies on an online music platform Last.fm show notable country- and age-based homophily [8, 9]
- Other results [8-12] show higher similarity in user music taste(based on artists or genres) for friends in comparison with random user pairs
- Graph-based features such as common friends are most indicative for friendship prediction [10]
- In Last.fm on average, communities identified with graph partitioning algorithms are not composed of users that listen to the same music [13]

[8] Bischoff, K. (2012, June). We love rock'n'roll: analyzing and predicting friendship links in Last. fm. In Proc. of the 4th Annual ACM WebSci Conference.
[9] Baym, N. K., & Ledbetter, A. (2009). Tunes that bind? Predicting friendship strength in a music-based social network. Information, Communication & Society.
[10] Aiello, L. M., Barrat, A., Schifanella, R., Cattuto, C., Markines, B., & Menczer, F. (2012). Friendship prediction and homophily in social media. TWEB.
[11] Zhou, Z., Xu, K., & Zhao, J. (2018). Homophily of music listening in online social networks of China. Social Networks.
[12] Guidotti, R., & Rossetti, G. (2019, October). “Know Thyself” How Personal Music Tastes Shape the Last. Fm Online Social Network. ISFM.
[13] Bisgin, H., Agarwal, N., & Xu, X. (2010, August). Does similarity breed connection?-an investigation in Blogcatalog and Last. fm communities. Social Computing.

Existing research

- The results of Guidotti et al. [12] show that Last.fm users who listen to various genres tend to connect with people with high music preference entropy (high diversity) and users who listen to music comprising few genres tend to connect with users with a narrow music taste
- A study on music platforms *Netease* and *Weibo* demonstrates that users with low diversity are more similar in terms of music taste and that it is difficult for high diversity users to find friends sharing similar music preferences [11]

[11] Zhou, Z., Xu, K., & Zhao, J. (2018). Homophily of music listening in online social networks of China. Social Networks.
[12] Guidotti, R., & Rossetti, G. (2019, October). “Know Thyself” How Personal Music Tastes Shape the Last. Fm Online Social Network. ISFM.

Research objective

*We investigate **homophily** and **link prediction** based on **user music taste**.*

*We approach this objective by investigating **homophily** and **link prediction** with users’ music preferences, such as **mainstreaminess**, **novelty**, **diversity**, and **artist profile similarity**.*

Research questions

I. *To which degree do users exhibit homophily with respect to features describing their preferences towards mainstream, diverse, or novel content?*

II. *How does it relate to homophily based on artists they listen to?*

III. *What is the merit of using such features for friendship prediction?*

Data and preprocessing

Dataset

- We use the well-established LFM-1b [13] and LFM-1b UGP [14] datasets which include the following data of our interest:
  i. User-artist playcount matrix $\boldsymbol{L}^{artist}$ (3,190,371 unique artists)
  ii. User-genre playcount matrix $\boldsymbol{L}^{genre}$(1,998 unique genres and styles from Freebase [15])
  iii. Features describing user mainstreaminess, novelty, and diversity
- LFM-1b contains more than one billion listening events (LEs) of 120,175 unique users of the Last.fm music platform

[13] Schedl, M. (2016, June). The lfm-1b dataset for music retrieval and recommendation.ICMR.
[14] Schedl, M., & Ferwerda, B. (2017, December). Large-scale analysis of group-specific music genre taste from collaborative tags. ISM.
[15] https://developers.google.com/freebase

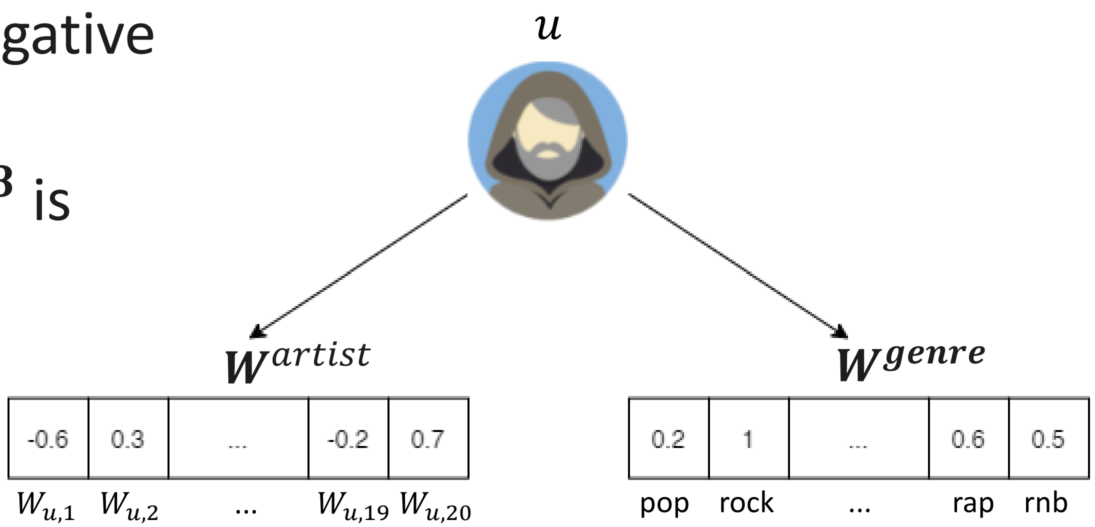
Friendship network

- Last.fm allows users to connect with others on the platform by establishing friendship connections
- A list of each user's friends is available through the Last.fm API [16]
- We adopt the breadth-first-search sampling strategy with LFM-1b users as seed nodes and crawl friends up to two hops away.
    - We also keep only those edges where both nodes are part of the LFM-1b dataset
- This results in an undirected unweighted graph $G = (V, E)$ consisting of $|V| = 11{,}792$ nodes and $|E| = 78{,}989$ edges (friendship connections)

[16] https://www.last.fm/api/show/user.getFriends

Listening profiles

- We create an artist and a genre profile for each user
- Artist profile matrix $\boldsymbol{W}^{artist} \in \mathbb{R}^{11,\mathbf{792}\times\mathbf{20}}$ is created by compressing $\boldsymbol{L}^{artist}$ using non-negative matrix factorization
- Genre profile matrix $\boldsymbol{W}^{genre} \in \mathbb{R}^{11,\mathbf{792}\times\mathbf{1,998}}$ is created by row normalizing $\boldsymbol{L}^{genre}$



User mainstreaminess (M)

- Mainstreaminess describes each user in terms of the degree to which they prefer music items that are currently popular or rather ignore such trends
- The general idea behind calculating M is to relate user playcounts to the global playcounts
- We adopt a number of M features from the LFM-1b dataset, calculated over time windows of 1, 6, and 12 months, as well as over the entire period of user's listening activity (M^{1m}/ M^{6m} / M^{12m} / M^{G})

User novelty (N)

- Novelty models the inclination of user u to listen to unknown music and is quantified by the percentage of new artists listened to, averaged over a time window
- We adopt three N features from the LFM-1b dataset, averaged over time windows of 1, 6, and 12 months (N^{1m}/ N^{6m} / N^{12m})

User diversity (D)

- The LFM-1b dataset does not include explicit features of diversity other than simple numbers of unique tracks and artists listened to by the user, we log normalize those counts and denote them as
- We also compute genre coverage and genre entropy (similarly as in [17]) using the genre profile as input
- Additionally, we propose a novel diversity feature, i.e., the *weighted average genre diversity* D^{w_avg}, calculated from the user's genre playcount vector as follows:

$$D_u^{w\_avg} = \frac{\sum_{i=1}^{n_{genre}} \frac{L_{u,i}}{\mathbf{max}\ (L_{u,*})}}{n_{genre}}$$

[17] Deldjoo, Y., & Schedl, M. (2019, September). Retrieving relevant and diverse movie clips using the mfvcd-7k multifaceted video clip dataset. CBMI.

User groups

- For each M, N, and D feature, we categorize users into a *low*, *mid*, or *high* group based on the corresponding feature value. Decision thresholds are calculated as in [7]:
  1. Feature values are first sorted in ascending order
  2. Feature values are then summed starting from the smallest value until a third of the total sum is reached, all of those users are assigned into the *low* value group
  3. Users are assigned similarly into the *medium* value group
  4. The rest are assigned into the *high* value group

[7] M. Schedl and D. Hauger (2015), Tailoring music recommendations to users by considering diversity, mainstreaminess, and novelty. SIGIR.

Preparing the dataset for link prediction

- We approach link prediction as a binary classification problem, i.e., predict a class label $y \in \{0, 1\}$ given a feature vector $\boldsymbol{x}$
- Due to extreme class imbalance, we adhere to the guideline in [] and randomly sample negative $|E|$ class instances (missing edges) so that the resulting dataset consists of $|E|$ examples
- To further account for randomness, we create 10 datasets in the same manner with different random seeds and average the results in our experiments over all 10 datasets

[18] Yang, Y., Lichtenwalter, R. N., & Chawla, N. V. (2015). Evaluating link prediction methods. Knowledge and Information Systems.

Preparing the dataset for link prediction

- To create the feature vector $\boldsymbol{x}$, we concatenate feature vectors $\boldsymbol{x}_u$, $\boldsymbol{x}_v$, and $\boldsymbol{x}_{\boldsymbol{\Delta}} = \boldsymbol{f}(\boldsymbol{x}_u, \boldsymbol{x}_v)$ which represents features derived from both users
- We categorize features into:
  1. M, N, and D features (MNDF)
  2. Artist profile features (AF)
  3. Graph-based features (GF)
     - Common neighbors, Jaccard index, and Adamic-Adar index

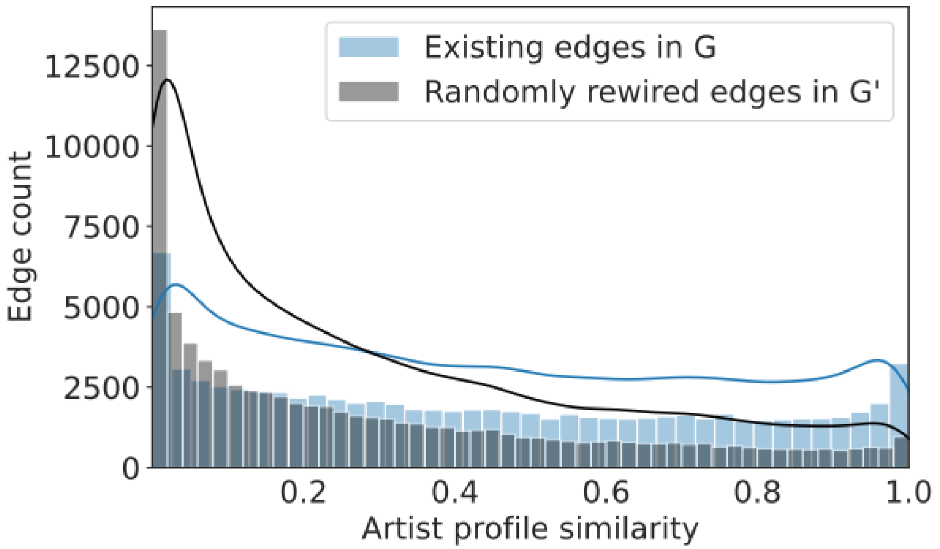
Resulting feature vector $\boldsymbol{x}$

Notation	Description	Type
	Mainstreaminess, novelty, and diversity features (MNDF)	
$\mathbf{x}^{\mathbf{M}}_{\{\mathbf{u},\mathbf{v}\}}$	4 mainstreaminess values per user	Numeric vector
$\mathbf{x}^{\mathbf{N}}_{\{\mathbf{u},\mathbf{v}\}}$	3 novelty values per user	Numeric vector
$\mathbf{x}^{\mathbf{D}}_{\{\mathbf{u},\mathbf{v}\}}$	5 diversity values per user	Numeric vector
$\mathbf{x}^{\mathbf{M_{group}}}_{\{\mathbf{u},\mathbf{v}\}}$	Low/mid/high group categorization for each mainstreaminess feature and each user	Categorical vector
$\mathbf{x}^{\mathbf{N_{group}}}_{\{\mathbf{u},\mathbf{v}\}}$	Low/mid/high group categorization for each novelty feature and each user	Categorical vector
$\mathbf{x}^{\mathbf{D_{group}}}_{\{\mathbf{u},\mathbf{v}\}}$	Low/mid/high group categorization for each diversity feature and each user	Categorical vector
$\mathbf{x}^{\mathbf{M}}_{\mathbf{\Delta}}$	Relative difference between respective user mainstreaminess features	Numeric vector
$\mathbf{x}^{\mathbf{N}}_{\mathbf{\Delta}}$	Relative difference between respective user novelty features	Numeric vector
$\mathbf{x}^{\mathbf{D}}_{\mathbf{\Delta}}$	Relative difference between respective user diversity features	Numeric vector
$\mathbf{x}^{\mathbf{M_{group}}}_{\mathbf{\Delta}}$	True if both users are categorized in the same mainstreaminess group, False otherwise	Binary vector
$\mathbf{x}^{\mathbf{N_{group}}}_{\mathbf{\Delta}}$	True if both users are categorized in the same novelty group, False otherwise	Binary vector
$\mathbf{x}^{\mathbf{D_{group}}}_{\mathbf{\Delta}}$	True if both users are categorized in the same diversity group, False otherwise	Binary vector
	Artist profile features (APF)	
$\mathbf{x}^{\mathbf{W^{artist}}}_{\{u,v\}}$	Low-dimensional user artist profile vectors	Numeric vector
$x^{\mathbf{W^{artist}}}_{\Delta}$	Cosine similarity between user artist profile vectors	Numeric scalar
	Graph-based features (GF)	
x^{CN}_{Δ}	Number of common neighbors between users	Numeric scalar
x^{J}_{Δ}	Jaccard index between users	Numeric scalar
x^{AA}_{Δ}	Adamic-Adar index between users	Numeric scalar

Homophily in user preferences

Artist profile homophily

- We investigate if ther is homophily in the Last.fm friendship graph G based on user's artist profiles by comparing cosine similarities between connected users and random user pairs from G'
- We create G' by randomly rewiring edges using an adapted configuration model which preserves the degree distribution and homophily based on M, N, and D user groups



Homophily based on M, N, and D

- Next, we quantify homophily in the Last.fm friendship network base on numeric M, N, and D feautures using the assortativity coefficient r:

$$r = \frac{\sum_{(u,v)\in E}(f(u) - \bar{f}(u))(f(v) - \bar{f}(v))}{\sqrt{\sum_{(u,v)\in E}(f(u) - \bar{f}(u))}\sqrt{\sum_{(u,v)\in E}(f(v) - \bar{f}(v))}}$$

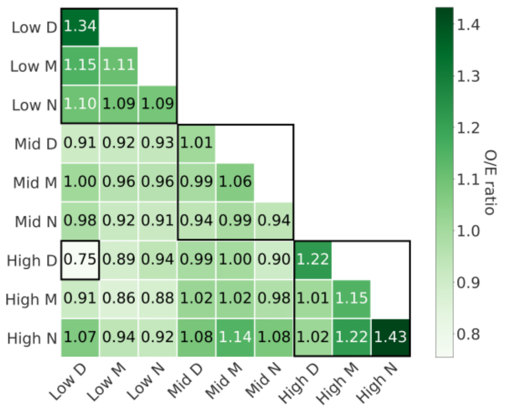
- We use the features with highest r as representative features in further experiments

		r
M	M^{1m}	0.055
	M^{6m}	0.050
	M^{12m}	0.043
	M^{G}	**0.104**
N	N^{1m}	0.063
	N^{6m}	**0.111**
	N^{12m}	0.058
D	D^{tracks}	0.076
	$D^{artists}$	0.104
	D^{GC}	0.148
	D^{GE}	0.140
	D^{w_avg}	**0.227**

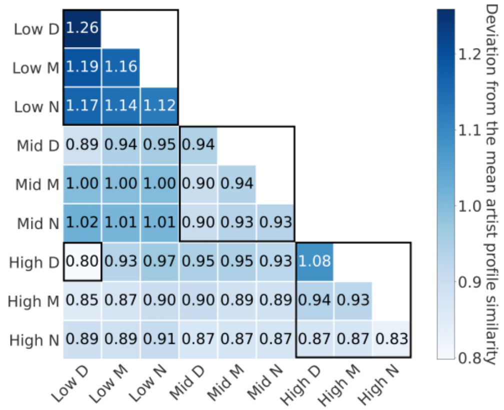
Between- and within-group observed to expected edge ratio

- Assortativity coefficients can only provides us with information on the overall assortativity patterns in the network
- However, what often happens is that homophily exists only on some intervals of the feature distribution, e.g., for users with particularly high or low values
- Using defined low/mid/high user groups, we investigate whether there is homophily in these user groups by counting the observed edges between or within groups and dividing it with the number of expected edges assuming fully random pairing
    - In a random graph, the expected number of edges within a particular group (e.g., low D) is $p^2 d(G)$ and between two different groups (e.g., between low D and high D) $2pqd(G)$, where p is the number of nodes in the first group, q is the number of nodes in the second group, and $d(G)$ is the density of the graph G
- O/E ratio within a group higher than 1 points to homophily
- To draw further insights, we also compare artist profile similarities between and within groups

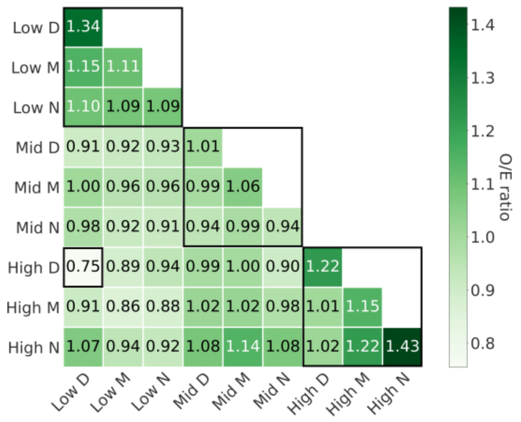
Between- and within-group observed to expected edge ratio



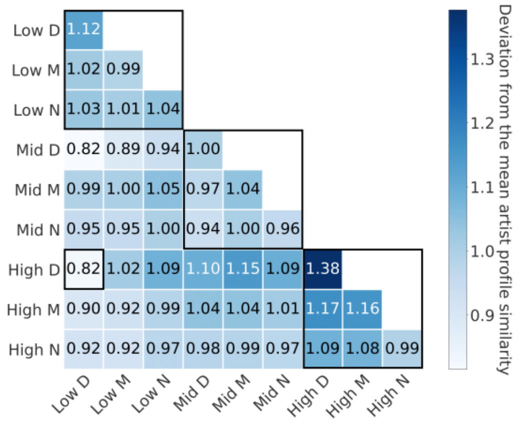
(a) O/E ratio in **G**



(b) Artist profile similarity in **G**



(c) O/E ratio in $\mathbf{G}'$



(d) Artist profile similarity in $\mathbf{G}'$

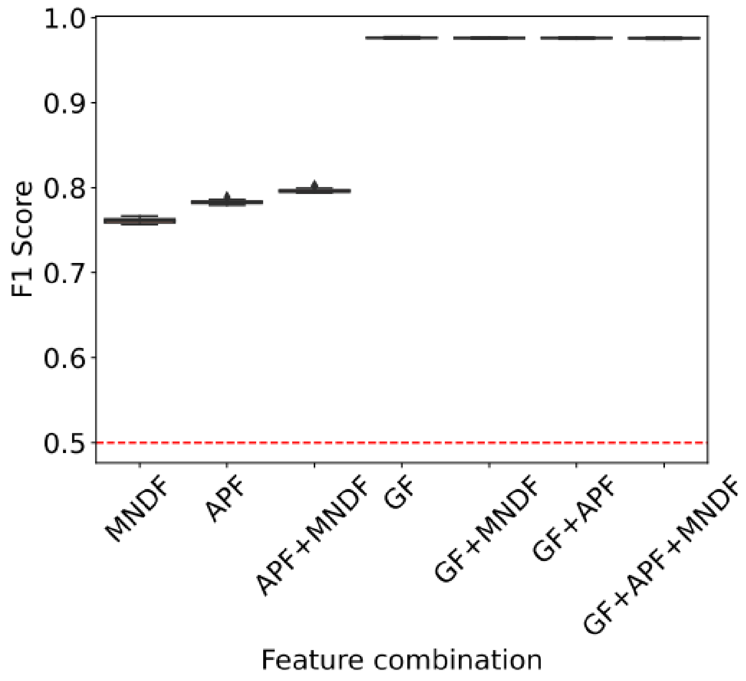
Link prediction with user preferences

Link prediction with user preferences (objectives)

- We don’t focus on performance comparison of different classification algorithms, but rather on using an established classification algorithm (XGBoost) [19] as a tool to explore the merit of M, N, and D features (**MNDF**) in comparison with user artist profile features (**APF**) for the task of link prediction
- We provide context to those results by comparing them with a strong baseline using graph-based features (**GF**) as well as a weak random baseline

[19] Behera, D. K., Das, M., Swetanisha, S., Nayak, J., Vimal, S., & Naik, B. (2021). Follower Link Prediction Using the XGBoost Classification Model with Multiple Graph Features. Wireless Personal Communications.

Accuracy results

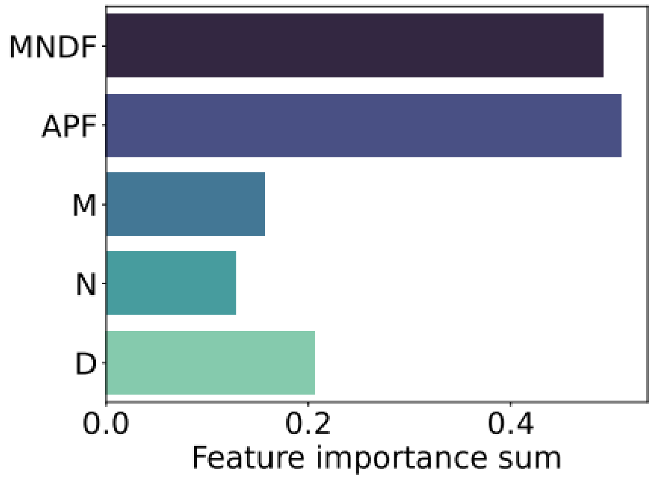


- Using MNDF outperforms the random baseline and achieves similar results to those achieved using APF, a combination both yields the best accuracy
- Both MNDF and APF are not able to outperform a strong baseline using GF for link prediction
  - Nevertheless, they can be useful in cold-start scenarios for user recommendation
- We also find notable differences in link prediction accuracy for different low/mid/high user groups

		MNDF	**APF**	**MNDF+APF**
M	Low	0.7668	0.7813	**0.7974**
	Medium	0.7602	0.7818	**0.7859**
	High	0.7761	**0.7946**	0.7908
N	Low	0.7655	0.7811	**0.7953**
	Medium	0.7643	0.7815	**0.7872**
	High	0.7811	0.7958	**0.7971**
D	Low	0.7671	0.7817	**0.7971**
	Medium	0.7602	0.7878	**0.7908**
	High	0.7668	0.7829	**0.7924**

Feature importance

- Furthermore, we explore the merit of using MNDF with APF for link prediction by looking into XGBoost feature importance scores for the MNDF+APF approach
- Out of single features, the one with the highest importance score is the artist profile similarity
  - This is followed by diversity features
- We also aggregate the feature importance scores into categories and show their sums
- We observe that MNDF and APF contribute almost equally to the final results of link prediction and D has more impact than M or N



Feature	Importance	Feature	Importance
$\mathbf{x}^{\mathrm{M}}_{\{u,v\}}$	0.0777	$\mathbf{x}^{\mathrm{N}}_{\Delta}$	0.0096
$\mathbf{x}^{\mathrm{N}}_{\{u,v\}}$	0.0600	$\mathbf{x}^{\mathrm{D}}_{\Delta}$	0.0180
$\mathbf{x}^{\mathrm{D}}_{\{u,v\}}$	0.1001	$\mathbf{x}^{\mathrm{M_{group}}}_{\Delta}$	0.0100
$\mathbf{x}^{\mathrm{M_{group}}}_{\{u,v\}}$	0.0559	$\mathbf{x}^{\mathrm{N_{group}}}_{\Delta}$	0.0094
$\mathbf{x}^{\mathrm{N_{group}}}_{\{u,v\}}$	0.0490	$\mathbf{x}^{\mathrm{D_{group}}}_{\Delta}$	0.0157
$\mathbf{x}^{\mathrm{D_{group}}}_{\{u,v\}}$	0.0726	$\mathbf{x}^{\mathrm{W_{artist}}}_{\{u,v\}}$	0.4130
$\mathbf{x}^{\mathrm{M}}_{\Delta}$	0.0124	$x^{\mathrm{W_{artist}}}_{\Delta}$	0.0964

Summary and conclusion

Summary and conclusion

- In this paper, we study homophily in an online music platform Last.fm regarding user preferences towards listening to mainstream (M), novel (N), or diverse (D) content
- Furthermore, we draw comparisons with homophily based on listening profiles derived from artists users have listened to, i.e., artist profiles
- Finally, we explore the utility of users’ artist profiles as well as features describing

Summary and conclusion

- Our findings show that users with a friendship connection share similar music taste based on their artist profiles
- Homophily is stronger for D, than for M and N
- Some user groups such as high-novelty-seekers (explorers) exhibit strong homophily, but lower than average artist profile similarity
- Using MNDF achieves comparable results on link prediction accuracy compared with using APF, but the combination of features yields the best accuracy results
- Using MNDF+APF does not add value if graph-based features such as common neighbors are available, making MNDF primarily useful in a cold-start user recommendation setting for users with few friendship connections

Thank you! Questions? Contact/follow us.

Get the paper:

https://arxiv.org/pdf/2111.00562

Social Computing Lab (@Know-Center and @ISDS)

know-center.tugraz.at/research/areas/social-computing/

s-computing@know-center.at

@socialcomplab

Tomislav Duricic

tduricic@tugraz.at

@djurazzi

scholar.google.com/citations?user=_mY5j2UAAAAJ

researchgate.net/profile/Tomislav_Duricic

linkedin.com/in/tomislav-đuričić-954b54172