Oh, Behave!
Country Representation Dynamics Created by Feedback Loops in Music Recommender Systems

Oleg Lesota[1], Jonas Geiger[1], Max Walder[1],
Dominik Kowald[2], Markus Schedl[1]

[1]Johannes Kepler University Linz and Linz Institute of Technology, [2]Know-Center GmbH and Graz University of Technology (Austria)

Motivation

- Previous work shows that recommender systems may under-represent music from certain countries in their recommendations.
- This may cause users receive less recommendations of their local (domestic) music and shift their preference towards global trends.
- This work investigates potential long-term effects of such under-representation.

Research questions

- RQ 1: How do different recommender systems affect representation of local and US-produced music in recommendations and user profiles in the long term?

- RQ 2: How do effects of feedback loops vary across different countries? Do different recommender systems treat individual countries differently?

Method

Feedback loop simulation:

1. Split current input data into train (75%), validation (20%), test (5%).
2. (Re)train the model, optimize for nDCG@10, produce 10 novel recommendations per user.
3. Simulate consumption of one item per user (randomly, higher probability for higher ranking items).
4. Combine simulated interactions with the current input. Go to step1.

Evaluation

- Average proportions of local and US music in recommendations and simulated consumptions histories (overall and per country).
- Jensen-Shannon Divergence (JSD) between 3-bin track popularity [HighPop, MidPop, LowPop] and country of origin (user-specific) [local, US-produced, other] distributions (miscalibration).

Dataset

	Tracks	Track Interactions		Users	User
		Total	Average		Interactions
US	39,614	1,040,360	26.26	1,582	323,072
UK	15,522	422,225	27.20	823	171,469
DE	6,793	107,832	15.87	805	158,642
SE	4,519	107,491	23.79	320	60,993
CA	3,754	95,343	25.40	217	47,490
FR	2,800	56,241	20.09	254	52,850
AU	2,346	53,701	22.89	193	40,767
FI	2,260	45,709	20.23	420	78,819
NO	1,765	36,769	20.83	208	40,545
BR	2,236	35,964	16.08	1,064	205,093
NL	1,738	32,035	18.43	375	89,546
PL	1,709	27,116	15.87	1,040	195,296
RU	1,888	24,086	12.76	1,162	187,876
JP	1,796	21,818	12.15	101	14,411
IT	1,506	21,273	14.13	222	37,421
other	9,651	159,769	16.55	2,990	583,442
Total	99,897	2,287,732	22.90	11,776	2,287,732

- User listening activity on Last.fm in 2018-2019 (LFM-2b dataset sample).
- Artist country information from MusicBrainz.org.

Paper & code

Results RQ1

Representation of Local and US-produced Music in the Long Term

Right: Proportion of local and US-produced music in recommendations (Rec∗) and simulated user profiles (Prof∗) at iteration 100, compared to the respective proportions in the original user profiles (difference in %). Significant changes are marked with ∗ .

	Pop	ItemKNN	BPR	NeuMF	MultiVAE	LightGCN
Rec_{local}	−47.5*	+0.4	−6.6*	−39.7*	−21.2*	−3.5
Rec_{US}	+15.6*	+4.8*	+3.5*	+17.0*	+7.5*	+0.5
$Prof_{local}$	−22.5*	+2.1	−7.0*	−19.5*	−10.5*	−3.6
$Prof_{US}$	+19.5*	+2.2*	+3.7*	+7.6*	+2.6*	+0.8
JSD_{Prof}	0.13	0.12	0.08	0.10	0.09	0.08
$nDCG_1$	0.03	0.26	0.13	0.08	0.12	0.14

- The proportion of local music declines in the recommendations and resulting profiles for most algorithms. The proportion of US-produced music generally increases.
- Varied impact patterns: LightGCN – preserves proportions, least miscalibrated (JSD), ItemKNN – preserves proportions, most miscalibrated, NeuMF – distorts proportions, miscalibrated.

- Iteration 1: most models already show local-item proportion inconsistent with the initial profiles (dashed line, Right).
- Some models may converge or remain around the initial user profiles (ItemKNN, BPR, LightGCN).
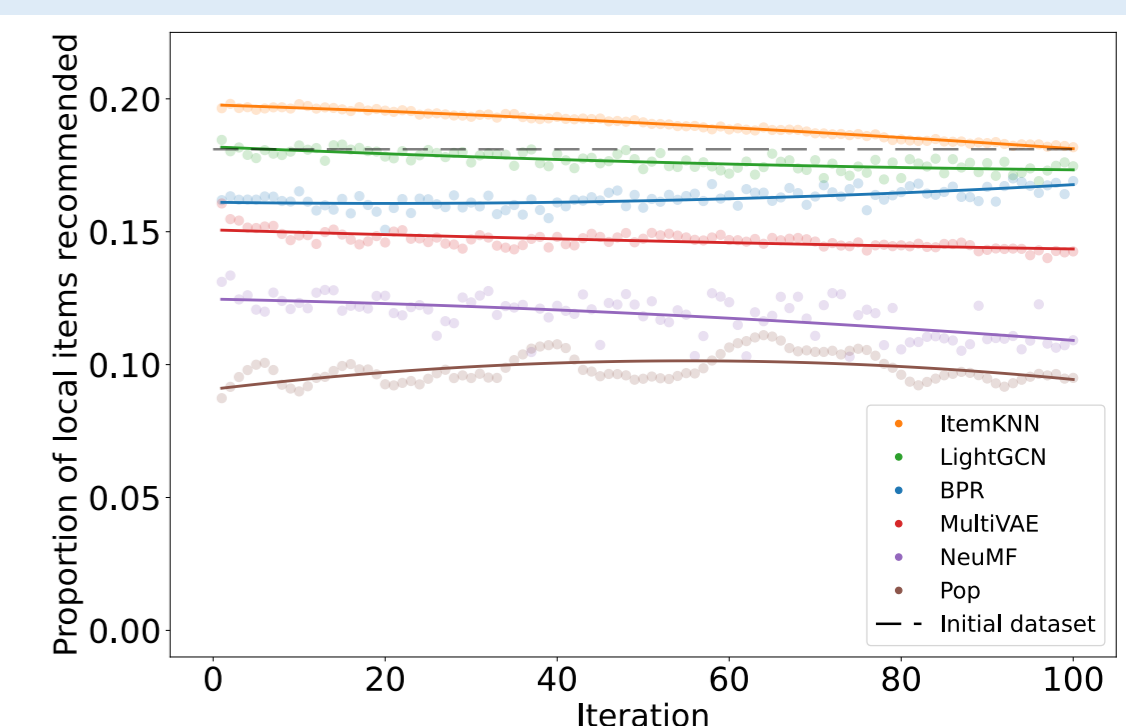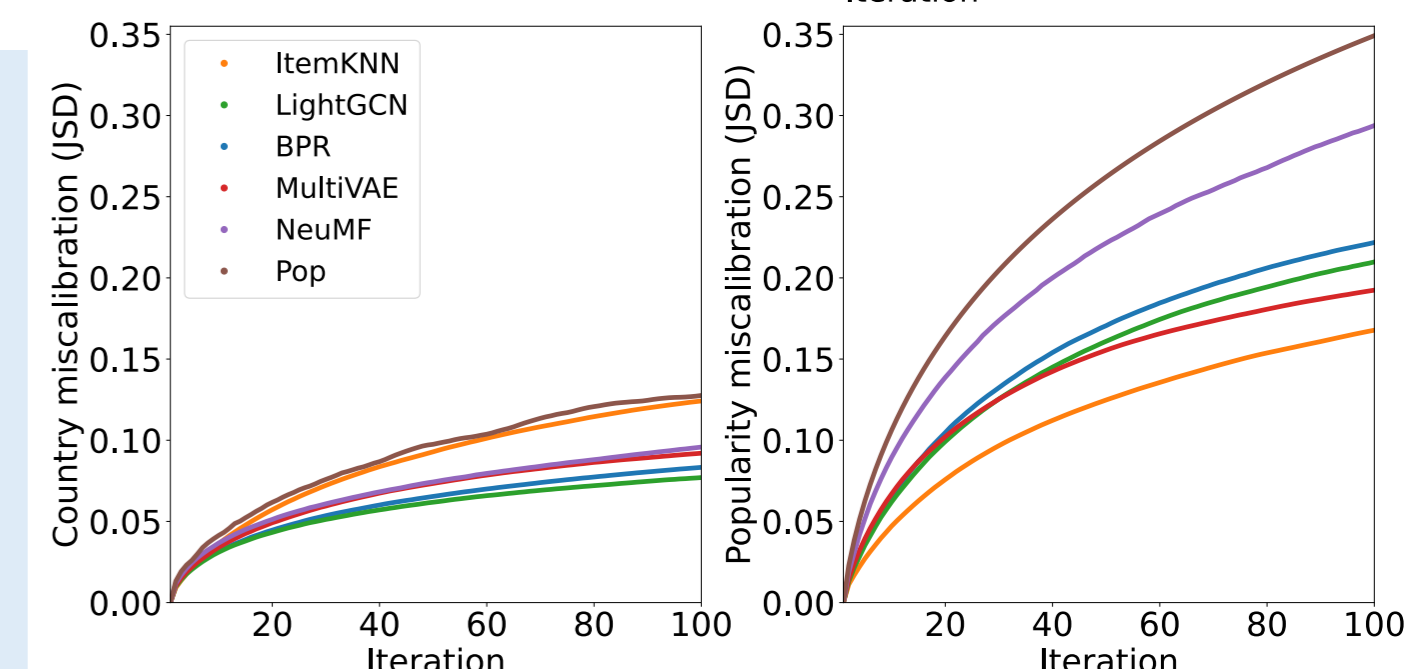

Note on calibration

ItemKNN:
- Highest country miscalibration
- Lowest popularity miscalibration

No direct connection between popularity and country calibration

Results RQ2

Effects of Feedback Loops Across Different Countries

Below: miscalibration between three-bin country distributions over the original user profiles and the profiles after 100 iterations (measured as JSD). Higher values indicate higher inconsistency.

	US	UK	DE	SE	CA	FR	AU	FI	NO	BR	NL	PL	RU	JP	IT	all
Pop	0.073	0.126	0.146	0.133	0.108	0.144	0.101	0.186	0.108	0.133	0.096	0.16	0.163	0.196	0.153	0.128
ItemKNN	0.084	0.123	0.139	0.137	0.106	0.128	0.109	0.152	0.127	0.134	0.094	0.155	0.149	0.166	0.174	0.124
BPR	0.054	0.093	0.089	0.09	0.085	0.095	0.088	0.103	0.084	0.083	0.07	0.103	0.098	0.146	0.114	0.083
NeuMF	0.058	0.099	0.114	0.104	0.09	0.111	0.093	0.126	0.087	0.099	0.074	0.121	0.115	0.179	0.135	0.096
MultiVAE	0.062	0.098	0.101	0.099	0.092	0.1	0.094	0.114	0.091	0.1	0.074	0.111	0.108	0.132	0.136	0.092
LightGCN	0.053	0.089	0.079	0.081	0.08	0.086	0.091	0.088	0.084	0.069	0.068	0.093	0.086	0.099	0.13	0.077

Below: deviations in the average proportions of local and US music in user profiles at iteration 100 from the respective average proportions in the original user profiles before simulation (in %). Statistically significant deviations are marked with *. Per column: highest value in **bold**, lowest value underlined.

	US	UK	DE	SE	CA	FR	AU	FI	NO	BR	NL	PL	RU	JP	IT	all
					local proportion in user profiles											
Pop	−1.2	−4.2	−41.6*	−40.8*	−31.0*	−42.2*	−31.2*	−53.4*	−47.3*	−48.6*	−44.1*	−54.4*	−57.7*	−67.3*	−58.3*	−22.5*
ItemKNN	**−1.5**	−15.0*	+5.6	+2.4	−18.3	−20.1	−25.6	**+16.0**	−1.7	+7.2	−19.6	**+28.5***	+10.7	+18.7	**+49.6**	**+2.1**
BPR	−2.3	−7.6*	−8.8	−17.9*	−23.7*	−31.3*	−31.3*	+3.4	−39.2*	+3.8	−35.1*	−1.2	−19.1*	−40.3	−5.6	−7.0*
NeuMF	−3.3*	**−6.1**	−33.6*	−31.4*	−23.1*	−40.2*	−33.8*	−35.0*	−39.9*	−29.5*	−37.3*	−46.7*	−49.0*	−66.5*	−56.6*	−19.5*
MultiVAE	−3.5*	−7.5*	−7.0	−19.0*	−23.4*	−34.8*	−29.8*	−19.0	−38.7*	−0.4	−36.3*	−23.6*	−12.7	−18.5	−53.1*	−10.5*
LightGCN	−3.1	−10.3*	−3.4	−1.7	−21.3*	**−3.7**	−29.5*	+6.9	−3.5	−31.5*	+12.3	−12.3	+10.0	**+48.6***	−3.6
					US proportion in user profiles											
Pop	−1.2	+14.9*	+28.8*	+23.8*	+2.3	+30.9*	+8.4*	+44.5*	+17.5*	+17.4*	+16.5*	+34.8*	+36.9*	+28.8*	+24.1*	+19.5*
ItemKNN	**−1.5**	**+8.4***	+2.9	−0.7	−0.4	+7.7	**+4.5**	−6.3	+1.5	−0.4	+5.7	+3.8	+6.1	−10.8	−1.9	+2.2*
BPR	−2.3	+7.3*	+4.3	+6.0	**+0.3**	+9.8	+3.2	+0.0	**+7.0**	+2.5	+4.8	+6.9*	+7.0*	−0.4	+3.1	+3.7*
NeuMF	−3.3*	+8.3*	**+13.6***	**+10.6***	−1.2	+15.4*	+2.9	+15.6*	+6.5	+9.7*	+6.7	+14.2*	+11.8*	+17.3*	+13.0*	+7.6*
MultiVAE	−3.5*	+5.2	+2.4	+3.4	+0.1	+6.0	+2.6	+4.4	+5.4	+2.8	+4.4	+6.4*	+3.0	+10.0*	+11.0*	+2.6*
LightGCN	−3.1	+5.2	+0.7	+0.2	−0.9	+3.0	+2.7	−2.6	+0.4	+1.9	+3.4	+3.2	+2.9	−5.6	−8.2	+0.8

- The list of significantly affected countries varies between algorithms. Overall, on per-country level most changes are in local rather than US proportions.
- Users from countries less represented in the data (e.g., FI) are more likely to receive recommendations inconsistent with their original preferences. However, higher representation does not always guarantee the opposite (e.g., DE).
- Causes of miscalibration and under-representation call for more research.

oleg.lesota@jku.at

https://github.com/hcai-mms/FeedbackLoops4RecSys