From The Beatles to Billie Eilish: Connecting Provider Representativeness and Exposure in Session-based Recommender Systems

Alejandro Ariza[1][0000−0002−3388−2316]⋆, Francesco Fabbri[2,3][0000−0002−9631−1799], Ludovico Boratto[2][0000−0002−6053−3015], and Maria Salamó[1][0000−0003−1939−896]

[1] Universitat de Barcelona, Barcelona, Spain
[2] Eurecat - Centre Tecnológic de Catalunya, Barcelona, Spain
[3] Pompeu Fabra University, Barcelona, Spain
alejandro.ariza14@ub.edu, francesco.fabbri@eurecat.org, ludovico.boratto@acm.org, maria.salamo@ub.edu

Abstract. Session-based recommender systems consider the evolution of user preferences in browsing sessions. Existing studies suggest as next item the one that keeps the user engaged as long as possible. This point of view does not account for the providers' perspective. In this paper, we highlight side effects over the providers caused by state-of-the-art models. We focus on the music domain and study how artists' exposure in the recommendation lists is affected by the input data structure, where different session lengths are explored. We consider four session-based systems on three types of datasets, with long, short, and mixed playlist length. We provide measures to characterize disparate treatment between the artists, through a systematic analysis by comparing (i) the exposure received by an artist in the recommendations and (ii) their input representation in the data. Results show that artists for which we can observe a lot of interactions, but offering less items, are mistreated in terms of exposure. Moreover, we show how input data structure may impact the algorithms' effectiveness, possibly due to preference-shift phenomena.

Keywords: Session-based recommender systems · Provider exposure.

1 Introduction

Recommender systems (RS) are key tools to support users in online platforms [16]. Recent literature has focused on monitoring the users in their browsing sessions, to generate adaptive recommendations in so called *session-based RS* [14]. Instead of considering only the historical interactions between users and items, session-based systems adapt in real time to user preferences.

While session-based systems focus on user effectiveness as their main goal, recently a multi-stakeholder perspective has become central, for both recommender and ranking systems [1,17]. RS can support this paradigm and consider

⋆ The first two authors contributed equally to this work.

providers' needs, by giving them a certain *exposure* when their items are recommended. However, recommendation technologies do not consider the provider perspective, thus overexposing popular providers [7, 13], often leading to unfair outcomes [5, 12]. In addition, the exposure in a ranking does not always match the expected one [4, 15]. Despite the growing interest on fairness in recommendation, session-based RS received less attention [7] and no study tackled the exposure generated by a given data distribution.

Contribution. In this work, we analyze how the input data distribution impacts over RS quality, focusing also on the final exposure given to providers. As use-case, we consider the music streaming scenario, considering data coming from user-song interactions in Last-FM [18]. We sample three datasets, characterized by short, long, and mixed session lengths. Inspired by recent studies comparing the effectiveness of neural and non-neural approaches [11], we also focus on these two classes, considering four session-based systems, two for each class. In our study, we go beyond provider popularity, trying to understand if the *representation* of an artist (i.e., how many items they have in their catalog) affects the exposure they are given. Our results show that size of input representation plays an important role, with big providers in terms of representation (e.g., number of items in the catalog) being exposed not only more than unpopular ones, but also more than *popular-but-smaller* ones. We quantify this effect showing a systematic bias against providers having less items, which get lower chances of being recommended, despite being very popular. In other words, new but very popular artists like Billie Eilish, with billions of streams in music platforms, would be recommended less than very popular but bigger acts in terms of representation, such as The Beatles.

In a summary, (*i*) we characterize the effectiveness of session-based RS, comparing different algorithms and datasets, (*ii*) we provide a measure of *expected exposure* and characterize its connection with *provider representativeness* and *relevance*, (*iii*) we delve into the causes behind disparate exposure.

2 Metrics and Algorithms

Nowadays, streaming music services process user-item interactions as time-framed sequences, known as *sessions*. Considering a session s_n as an ordered list of user-item interactions of length n, a RS tries to predict the interaction i_{n+1} at time $n+1$, suggesting a top-k list of most likely future interactions.

Performance assessment. In addition to traditional metrics, such as precision (P@K), recall (R@K), and mean average precision (MAP@K), metrics such as mean reciprocal rank (MRR@K) and hit rate (HR@K) have been introduced to focus only on the single highest-ranked relevant item [8, 14]. These metrics optimize model performances in terms of user preferences, without accounting for the other stakeholders, such as the item providers. For this reason, we introduce a metric to quantify the goodness of the tested models w.r.t. the artist's utility.

Provider Exposure. Provider exposure assesses the quality of the models from the perspective of the searched/recommended individuals [19]. We consider each session s_n of length n as a query, $q(s_n)$, submitted to the RS; each query is processed by the recommendation algorithm that returns a top-k list of items L, ordered by interaction probability. Hence, we can define the probability distribution of interactions as $\sum_{i \in \mathcal{I}} p(i|q(s_n))$, with $\mathcal{I}$ as the set of items, and $p(i|q(s_n))$ as the probability that the user will interact with the item i, defined as:

$$p(i|q(s_n)) = \frac{1/\log_2(pos_i + 1)}{\sum_{j \in L} 1/\log_2(pos_j + 1)}$$

Where pos_j is the position of the item j in the list L. After processing a relevant number of queries $\mathcal{Q}$, it is possible to aggregate all the probabilities involving the item i, defining the related *expected exposure*:

$$e_i(\mathcal{Q}) = \sum_{q \in \mathcal{Q}} p(i|q(s_n))$$

This measure is inspired by the one by Diaz et al. [19]; in presence of a relevant number of queries, it expresses the expected amount of interactions for an item.

Assuming to group items by providers, where $\mathcal{I}_p \subseteq \mathcal{I}$ is the subset of items sold by the provider p, we can define the *expected provider exposure* as:

$$e_p(\mathcal{Q}) = \sum_{i \in \mathcal{I}_p} \sum_{q \in \mathcal{Q}} p(i|q(s_n))$$

For brevity, since we consider the same set of queries for each dataset, we use e_p. The expected provider exposure can be compared with the one in the input data, indicated as e_p^*, which is the number of times items from a provider p have been selected within the test-set. In the next section, we explore how these new exposure measures differ, depending on different input data and $|\mathcal{I}_p|$.

3 Experiments

3.1 Data and Algorithms

We analyze listening events of the *last.fm* platform. The dataset contains 1B listening events, 32M items, and 3M providers [18]. Since listenings come with a timestamp, we can aggregate them in sessions, fixing a threshold to split them in ordered lists. Initial tests led us to choose 15 minutes as cut-off. We extract three samples from the dataset. In each case, we randomly sample 200k sessions and keep those with at least 3 listenings. We obtain the following datasets (details in Table 1): (*i*) **LFM-S** is composed by short sessions, with length in the range $[5, 25]$; (*ii*) **LFM-L** contains long sessions, with length in the range $[40, 200]$; (*iii*) **LFM-M** does not show differences in terms of session length.

As algorithms, we considered two neural and two non-neural approaches [11]. Association rules (**AR**), a non-neural one, considers co-occurrences at pairwise

Name	Events	|S|	|I|	Providers
LFM-S	1,087,808	154,452	148,591	18,464
LFM-L	4,846,552	95,672	477,991	46,310
LFM-M	2,451,790	171,341	278,195	30,311

Table 1. Summary of sampled listenings data with dataset name, number of listenings, number of distinct items and number of providers

level. The second non-neural approach is a nearest-neighbour algorithm at session level (**S-KNN**). One of the neural approaches is based on recurrent neural networks (**GRU4REC**) [9]. The other, (**NARM**) (supposedly an improvement of GRU4REC), uses attention mechanisms [10].The last 20% of the sessions of each dataset is used as test set and we generate top-20 lists. Hyperparameters are tuned as in the last benchmark paper [11].

3.2 Results

Algorithms' evaluation. We look at both accuracy metrics and our new exposure metrics. The distribution of the expected exposure (e_p) generated by the recommendations, is normalized by the real one (e_p^*). This metric is assumed to be constant and close to 1 in the best scenario, where the recommender is able to predict in the long run the exposure of each artist. Table 2 summarizes our findings. For each dataset and column, we indicate in bold the best model. The last two columns show the average of e_p/e_p^* and the relative standard deviation.

As the first three columns show, S-KNN is the most effective approach in all datasets, minus the long-session one (LFM-L), which shows slightly better MAP and R values with the AR algorithm. Our results confirm recent findings, with the neural-based approaches outperformed by the memory-based ones. Indeed, neural approaches are optimized to predict the next item. Surprisingly, also considering the metrics coherent with their neural approaches' optimisation (HR@20 and MRR@20), the neural approaches do not always outperform the other methods. When comparing the datasets, the short-session one (LFM-S) produces the most effective predictions. Hence, when sessions get longer, algorithms cannot capture users' interests and understand what might be relevant for them. These results can be better understood by considering the metrics referred to the ratio between expected and real exposure, in the last two columns. Long sessions present the worst disparate exposure, confirming the algorithms are not able to catch drifts in user interests along the session. This leads to unstable exposure along the providers, leading to the highest values for $\mu(e_p/e_p^*)$ and $\sigma(e_p/e_p^*)$. Another interesting phenomenon in the last two columns is that NARM returns a distribution of providers exposure closest to the test, thus creating a trade-off between recommendation effectiveness and distribution of providers.

Impact of Provider Representativeness. Since the last two columns in Table 2 showed a clear instability of the algorithms to connect consistently expected

Name	Algorithm	MAP@20	P@20	R@20	HR@20	MRR@20	$\mu(e_p/e_p^*)$	$\sigma(e_p/e_p^*)$
LFM-S $\|S\| = 154{,}452$ $\bar{s}_l = 7.04$	AR	0.0421	0.0848	0.3769	0.4630	0.1789	0.8907	0.6952
	S-KNN	**0.0446**	**0.0905**	**0.4110**	**0.5153**	0.1410	0.7679	**0.5787**
	GRU4Rec	0.0254	0.0588	0.2882	0.4328	**0.3262**	1.1792	1.4163
	NARM	0.0301	0.0680	0.3234	0.4505	0.2641	**0.8909**	0.9757
LFM-L $\|S\| = 95{,}672$ $\bar{s}_l = 50.66$	AR	**0.0243**	0.1418	**0.1332**	**0.3349**	0.0915	1.1913	2.4121
	S-KNN	0.0226	**0.1460**	0.1174	0.2747	0.0663	0.6277	**1.1023**
	GRU4Rec	0.0084	0.0672	0.0665	0.3130	**0.2292**	1.7210	16.9100
	NARM	0.0129	0.0976	0.0789	0.1936	0.0537	**1.0195**	4.2379
LFM-M $\|S\| = 171{,}341$ $\bar{s}_l = 14.31$	AR	0.0302	0.1098	0.2258	0.3743	0.1219	1.0840	3.5950
	S-KNN	**0.0339**	**0.1295**	**0.2481**	0.3974	0.1019	0.5953	**0.6412**
	GRU4Rec	0.0186	0.0796	0.1802	0.3770	**0.3262**	1.2481	1.8928
	NARM	0.0261	0.1064	0.2116	0.3740	0.1540	**0.9506**	4.6853

Table 2. Performance for four algorithms tested on three different datasets, in terms of accuracy and providers exposure.

artists' exposure with the ground truth, we investigate the possible sources of this effect. We look at the impact of the *provider representativeness* $\mathcal{I}_p$ and input relevance $rel_p = log_{10}(|\mathcal{E}_p|)$, where $\mathcal{E}_p$ is the number of events within the training data, which involve an item of a provider p. We generate, for each use-case, a scatter plot, where each point presents on the x-axis the logarithm of provider representativeness, $log_{10}(|\mathcal{I}_p|)$, and on the y-axis the ratio of expected and real exposure, $log_{10}(e_p/e_p^*)$. The dots are colored by the provider relevance rel_p and logarithmic scale is needed for the two axes, so that we can have an homogeneous representation, including also the possible outliers in the analysis. From Fig. 1, a common pattern emerges: artists with a higher value of $|\mathcal{I}_p|$, are also the most relevant. Interesting is also the fact that providers with bigger $|\mathcal{I}_p|$ (right side of the plots) present a fair value of e_p/e_p^* and are not overexposed. However, being a relevant provider, but not having many items in the market $|\mathcal{I}_p|$ (like emerging artists) may impact negatively on the e_p/e_p^* value. This means that a small *provider representativeness* affects the ability to return a fair value of e_p/e_p^* (i.e., in the plot, it is fair when close to 0). The left part of all the scatters shows how blurry are the sections of dots involving relevant and non-relevant artists, revealing how all the algorithms are unable to catch differences in relevance among artists having small $|\mathcal{I}_p|$. The neural approaches, which present higher $\sigma(e_p/e_p^*)$ in Table 2, confirm to be the most challenged. Among them, S-KNN is the most stable along the datasets and GRU4REC the worst.

Impact of Session Length. The three datasets, characterized by different ranges of session length, raise concerns on the limitations and common issues of state-of-the-art session-based algorithms. Longer-session data (LFM-L), reveals that longer sequences of interactions increase the unpredictability for the user, leading to a precarious artists representation. All the models present higher range of e_p/e_p^* if compared with the other two datasets. On the other hand, shorter-session data (LFM-S) helps the model to provide more stable recommendations, where representativeness is consistently decoupled from relevance in all the approaches.

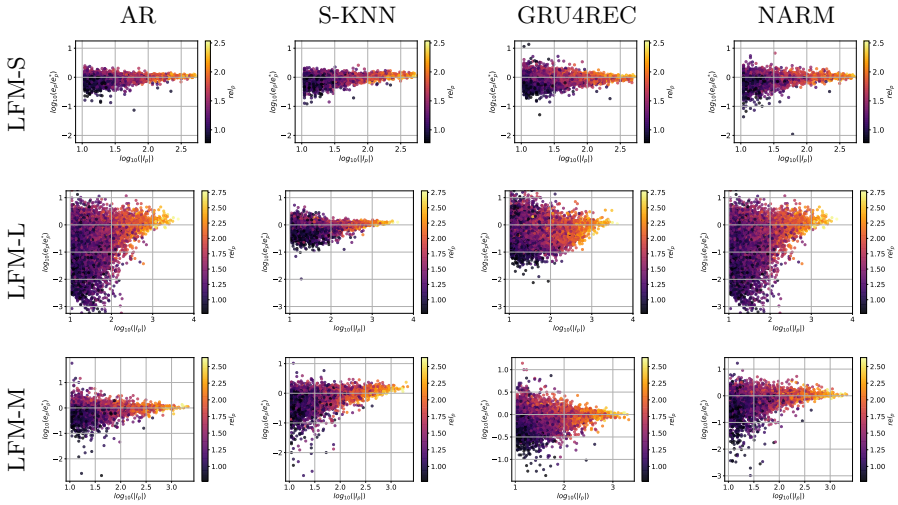

Fig. 1. Scatter plot, capturing the relationship between exposure, representativeness, and relevance of each provider. Over the x-axis the number of items produced by the provider p, over the y-axis the ratio between exposure by recommendations and one by test-set; each dot is colored with the relevance of the provider in the training set.

4 Conclusions

In this paper, we analyzed session-based RS, uncovering performance limitations due to different input data characteristics. Our findings align with recent work that sheds light on the limited progress of state-of-the-art models; in addition, we introduce the role of data distribution in this conversation. For the consumer-side, if we do not account for distribution of longer and shorter sessions, the effectiveness evaluation may be misleading. In addition, optimizing the accuracy leads to mistreatment and disparate exposure for providers. This finding connects our work to algorithmic fairness, by showing the incapability of models to calibrate the output, given the provider input relevance and representativeness. In the future, we will consider group-based scenarios, for both providers and consumers. We will also consider different datasets and session-based domains [2,3,6], with multiple definitions of exposure. From the algorithmic fairness perspective, we expect to design new session-based algorithms to meet exposure policies based on statistical parity or disparate treatment.

Acknowledgments. This research was partially funded by project 2017-SGR-341, MISMIS-LANGUAGE (grant No. PGC2018-096212-B-C33) from the Spanish Ministry of Science and Innovation, and NanoMoocs (grant No. COMRDI18-1-0010) from ACCIÓ. L. Boratto and F. Fabbri acknowledge ACCIÓ, for its support under project “Fair and Explainable Artificial Intelligence (FX-AI)”.

References

1. Abdollahpouri, H., Adomavicius, G., Burke, R., Guy, I., Jannach, D., Kamishima, T., Krasnodebski, J., Pizzato, L.A.: Beyond personalization: Research directions in multistakeholder recommendation. CoRR **abs/1905.01986** (2019)
2. Barra, S., Marras, M., Fenu, G.: Continuous authentication on smartphone by means of periocular and virtual keystroke. In: Au, M.H., Yiu, S., Li, J., Luo, X., Wang, C., Castiglione, A., Kluczniak, K. (eds.) Network and System Security - 12th International Conference, NSS 2018, Hong Kong, China, August 27-29, 2018, Proceedings. Lecture Notes in Computer Science, vol. 11058, pp. 212–220. Springer (2018). https://doi.org/10.1007/978-3-030-02744-5_16
3. Dessì, D., Fenu, G., Marras, M., Reforgiato Recupero, D.: COCO: semantic-enriched collection of online courses at scale with experimental use cases. In: Rocha, Á., Adeli, H., Reis, L.P., Costanzo, S. (eds.) Trends and Advances in Information Systems and Technologies - Volume 2 [WorldCIST'18, Naples, Italy, March 27-29, 2018]. Advances in Intelligent Systems and Computing, vol. 746, pp. 1386–1396. Springer (2018). https://doi.org/10.1007/978-3-319-77712-2_133
4. Diaz, F., Mitra, B., Ekstrand, M.D., Biega, A.J., Carterette, B.: Evaluating stochastic rankings with expected exposure. CoRR **abs/2004.13157** (2020)
5. Fenu, G., Lafhouli, H., Marras, M.: Exploring algorithmic fairness in deep speaker verification. In: Gervasi, O., Murgante, B., Misra, S., Garau, C., Blecic, I., Taniar, D., Apduhan, B.O., Rocha, A.M.A.C., Tarantino, E., Torre, C.M., Karaca, Y. (eds.) Computational Science and Its Applications - ICCSA 2020 - 20th International Conference, Cagliari, Italy, July 1-4, 2020, Proceedings, Part IV. Lecture Notes in Computer Science, vol. 12252, pp. 77–93. Springer (2020). https://doi.org/10.1007/978-3-030-58811-3_6
6. Fenu, G., Marras, M.: Leveraging continuous multi-modal authentication for access control in mobile cloud environments. In: Battiato, S., Farinella, G.M., Leo, M., Gallo, G. (eds.) New Trends in Image Analysis and Processing - ICIAP 2017 - ICIAP International Workshops, WBICV, SSPandBE, 3AS, RGBD, NIVAR, IWBAAS, and MADiMa 2017, Catania, Italy, September 11-15, 2017, Revised Selected Papers. Lecture Notes in Computer Science, vol. 10590, pp. 331–342. Springer (2017). https://doi.org/10.1007/978-3-319-70742-6_31
7. Ferraro, A., Jannach, D., Serra, X.: Exploring longitudinal effects of session-based recommendations. arXiv preprint arXiv:2008.07226 (2020)
8. Hidasi, B., Karatzoglou, A.: Recurrent neural networks with top-k gains for session-based recommendations. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. pp. 843–852. ACM (2018)
9. Hidasi, B., Karatzoglou, A.: Recurrent neural networks with top-k gains for session-based recommendations. In: Cuzzocrea, A., Allan, J., Paton, N.W., Srivastava, D., Agrawal, R., Broder, A.Z., Zaki, M.J., Candan, K.S., Labrinidis, A., Schuster, A., Wang, H. (eds.) Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018. pp. 843–852. ACM (2018). https://doi.org/10.1145/3269206.3271761
10. Li, J., Ren, P., Chen, Z., Ren, Z., Lian, T., Ma, J.: Neural attentive session-based recommendation. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. pp. 1419–1428 (2017)
11. Ludewig, M., Mauro, N., Latifi, S., Jannach, D.: Performance comparison of neural and non-neural approaches to session-based recommendation. In: Proceedings of the 13th International ACM RecSys Conference on Recommender Systems (2019)

12. Marras, M., Korus, P., Memon, N.D., Fenu, G.: Adversarial optimization for dictionary attacks on speaker verification. In: Kubin, G., Kacic, Z. (eds.) Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019. pp. 2913–2917. ISCA (2019). https://doi.org/10.21437/Interspeech.2019-2430
13. Mehrotra, R., McInerney, J., Bouchard, H., Lalmas, M., Diaz, F.: Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. pp. 2243–2251. ACM (2018)
14. Quadrana, M., Cremonesi, P., Jannach, D.: Sequence-aware recommender systems. ACM Comput. Surv. **51**(4), 66:1–66:36 (2018)
15. Ramos, G., Boratto, L., Caleiro, C.: On the negative impact of social influence in recommender systems: A study of bribery in collaborative hybrid algorithms. Inf. Process. Manag. **57**(2), 102058 (2020). https://doi.org/10.1016/j.ipm.2019.102058
16. Ricci, F., Rokach, L., Shapira, B., Kantor, P.B.: Recommender Systems Handbook. Springer-Verlag, Berlin, Heidelberg, 1st edn. (2010)
17. Saúde, J., Ramos, G., Caleiro, C., Kar, S.: Reputation-based ranking systems and their resistance to bribery. In: Raghavan, V., Aluru, S., Karypis, G., Miele, L., Wu, X. (eds.) 2017 IEEE International Conference on Data Mining, ICDM 2017, New Orleans, LA, USA, November 18-21, 2017. pp. 1063–1068. IEEE Computer Society (2017). https://doi.org/10.1109/ICDM.2017.139
18. Schedl, M.: The lfm-1b dataset for music retrieval and recommendation. In: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval. pp. 103–110 (2016)
19. Singh, A., Joachims, T.: Fairness of exposure in rankings. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 2219–2228. ACM (2018)