Title	Are You Reaching Your Audience? Exploring Item Exposure over Consumer Segments in Recommender Systems
Authors(s)	Wasilewski, Jacek, Hurley, Neil J.
Publication date	2017-08-31
Publication information	Wasilewski, Jacek, and Neil J. Hurley. "Are You Reaching Your Audience? Exploring Item Exposure over Consumer Segments in Recommender Systems." ACM, 2017.
Conference details	The 11th ACM Conference on Recommender Systems, Como, Italy, 27-31 August 2017
Publisher	ACM
Item record/more information	http://hdl.handle.net/10197/9979

Are You Reaching Your Audience? Exploring Item Exposure over Consumer Segments in Recommender Systems

Jacek Wasilewski
Insight Centre for Data Analytics
University College Dublin
Dublin, Ireland
jacek.wasilewski@insight-centre.org

Neil Hurley
Insight Centre for Data Analytics
University College Dublin
Dublin, Ireland
neil.hurley@insight-centre.org

ABSTRACT

Many state-of-the-art recommender systems are known to suffer from popularity bias, which means that they have a tendency to recommend items that are already popular, making those items even more popular. This results in the item catalogue being not fully utilised, which is far from ideal from the business' perspective. Issues of item exposure are actually more complex than simply overexposure of popular items. In this paper we look at the exposure of individual items to different groups of consumers, the item's *audience*, and address the question of whether recommender systems reach each item's potential audience. Thus, we go beyond state-of-the-art analyses that have simply addressed the extent to which items are recommended, regardless of whether they are reaching their target audience. We conduct an empirical study on the MovieLens 20M dataset showing that recommender systems do not fully utilise items' audiences, and existing sales diversity optimisers do not improve their exposure.

1 INTRODUCTION

Recommender systems have become ubiquitous in the interfaces to product catalogues provided by on-line retailers. From the user's perspective, recommender algorithms are used to filter a large set of possible selections into a much smaller set of items that the user is likely to be interested in. However, engaging and holding a user's interest is a complex matter and simply identifying relevant items might not be enough to satisfy a user. On the other hand, from the business point of view, as important as users receiving engaging recommendations, is the utilisation of products in the catalogue.

It has been shown [5] that on-line systems suffer from the *long tail effect*, where a few items (the "head") are responsible for the most of the interactions, while the rest (the "long tail") are much less known. Many of the state-of-the-art collaborative filtering recommender engines worsen the situation and promote already popular items, making both users and business suffer—only few items are engaged in sales and users are offered popular items of little surprise. Promoting the long tail items in recommendations offers benefits to both sides. It helps users to discover new items, which corresponds to the goal of a recommender system and, from the business side, it exposes the full product catalogue, increasing the chance of long-tail sales.

Viewed from the business perspective, recommender systems are marketing tools that identify which customers a product in the catalogue should be promoted to. Each item has its audience to which the item ideally should be exposed and making sure that items are exposed to the right users is an important step towards increasing sales. As one illustration, consider as an example the movie, *The Shining*, a well-known and acclaimed movie. Users may have watched it for various reasons, e.g. because it was directed by Stanley Kubrick, or because Jack Nicholson was the leading actor, or because it is based on a Stephen King novel, or because it is a horror, or just because it is a popular movie. Marketeers should ideally identify all these user interests and make sure the movie is exposed to users with similar interests in future, so that, if one of the movie's main attractions is the writer of the novel on which it is based, it is exposed to people who showed an interest in King's other novels.

Automating this item perspective in a top-N recommender system is not straight-forward. As each interaction is the recommendation of N relevant items to a single user, the system can directly control the user-perspective. In contrast, item exposure is built through a sequence of separate interactions with different users. Ensuring that an item is recommended to a relevant user, does not necessarily imply that this item's set of recommendations are properly spread across its full potential audience.

In this paper we tackle the problem of *item exposure* over consumer segments to understand who consumes items and if potential consumers are reached by recommendations. While there have been many studies of long-tail sales promotion, this work differs from previous studies and evaluation techniques in that it considers not only the number of times items are exposed but also to whom it is exposed, not treating all users the same. The main goal of this paper is to explore : a) whether items have specialised audiences, b) whether item audiences are fully targeted by recommender systems, and c) whether state-of-the-art sales diversity re-rankers improve exposure to audiences.

2 BACKGROUND

Improving user satisfaction and optimising business performance are two of the main goals that recommender systems are tasked with. While analysing different aspects of user satisfaction has drawn more attention in the research community, the business-oriented perspective on how recommender systems perform is equally important.

"Selling less of more"[5, 6], a concept formulated by Anderson suggests that the full catalogue of items should be utilised rather than having sales concentrated around a few heavily consumed items. Recommending less obvious items can potentially optimise both the user and business goals; the business promotes less popular but still relevant items, resulting in a richer user experience.

, ,
.

A number of researchers have looked at the problem of catalogue utilisation in recommender systems. Adomavicius and Kwon [3, 4] proposed a measure called *Aggregate Diversity*, defined as the total number of items that a system recommends to a given set of users. It measures the extent to which the item catalogue is exposed to the users of the system and can also be understood as a measure of item coverage. Following the “selling less of more” idea, systems with high Aggregate Diveristy are preferred. It has been shown, however, that classic collaborative filtering algorithms do not expose the full catalogue to the users. In fact they are biased towards recommending items coming from the so-called, *short head*—items that account for a significant portion of interaction with users—as opposed to the rest, the *long tail*. Celma and Herrera [7, 8] showed that the topology of the item similarity network could be the reason for poor discovery and low catalogue utilisation.

Another notion of item performance was presented by Fleder and Hosanagar in [10, 15], where sales diversity was measured through the sales concentration defined by the *Gini Index*. As opposed to the Aggregate Diversity, this metric does not only check whether items have been consumed, but how evenly/unevenly consumption is distributed across items. A similar measure of distributional inequality is the *Shannon Entropy* [15, 16] which is 0 if only one item is utilised by a system, and $\log|\mathcal{I}|$ if all items are recommended equally often. The Gini Index and Shannon Entropy measure pure distributional dispersion of item consumption or recommendation but do not account for prior item popularity. They cannot therefore be used to examine whether a recommender system reinforces or reduces prior concentration of popularity. Adamopoulos and Tuzhilin [1, 2] proposed popularity reinforcement measures to assess whether recommender systems follow prior popularity of items.

Several methods to promote sales diversity have been proposed [3, 4, 14, 17, 18, 20]. A simple approach is to promote items that are more likely to be unknown by a user, through expected popularity complement (EPC), defined as item novelty:

$$nov(i) = 1 - \frac{|\mathcal{U}_i|}{|\mathcal{U}|}$$

where $|\mathcal{U}_i|$ is item popularity and $|\mathcal{U}|$ is the total number of users. Recommendations can be re-scored by a linear combination of scores $s(u, i)$ provided by the recommendations and item novelty:

$$s^*(u, i) = (1 - \lambda)s(u, i) + \lambda nov(i)$$

λ is used to control the balance between relevance and novelty.

In all of the above, item exposure is measured by counting the number of times items are being consumed, i.e. item popularity. This assumes that all users consuming an item are alike and does not distinguish between users. In marketing, it is quite common to split the user-base into consumer segments in order to examine how each segment perceives an item of interest. Prior work leaves open the question of how items are exposed to different segments of people, and what is the impact of recommender systems on exposure across segments.

3 ITEM EXPOSURE OVER CONSUMER SEGMENTS

Recommender systems have to deal with the long tail of items that are rarely recommended. This includes niche items that are rarely liked, but also includes items that have not yet succeeded in penetrating the market. The interaction matrix may contain few entries associated with such items because users are not aware of their utility. Increasing their exposure may result in them eventually cascading through the marketplace and their popularity increasing. To identify and promote these items, we argue it is not enough to ask *how many* users have rated each item in the past, but also *which users* have rated the items, which define its *item exposure*.

By looking at the set of users who consumed and rated an item in the past, the item’s user profile $\mathcal{U}_i$, it is possible to model exposure of the item to different user types. Similarly, the set of users to whom the item is recommended, $\mathcal{R}_i$, can be analysed to reveal the extent to which a recommendation algorithm follows the exposure of the item. For that, a notion of user types is required.

As it is commonplace for marketeers to model their customer-base through customer segmentation, we find it useful to measure the exposure in terms of the spread across different consumer segments, where a segment represents a common taste. Given a partition of the taste space C into k consumer segments, we define *item exposure* of a set of consumers $\mathcal{S}$ as a function of the probability distribution P over consumer segments, where probability $P(c)$ describes the consumers’ preference towards segment c, and $\sum_{c \in C} P(c) = 1$. To measure the item exposure, we find it useful to compare the distribution P against a baseline distribution Q, which we consider to be one of ideal exposure. Then, a useful measure of the exposure of P, is its distance from the ideal distribution. We can measure the distance in terms of the Kullback-Leibler (KL) divergence:

$$D_{\mathrm{KL}}(P||Q) = \sum_{c \in C} P(c) \log_2 \frac{P(c)}{Q(c)}$$

which can be seen as the relative entropy of P with respect to Q. It is easier, however, to interpret another divergence, based on KL, the Jensen-Shannon (JS) divergence, as it is symmetric, always finite and bounded between 0 and 1 if $\log_2$ is used in D_{KL}:

$$D_{\mathrm{JS}}(P||Q) = \frac{1}{2} D_{\mathrm{KL}}(P||M) + \frac{1}{2} D_{\mathrm{KL}}(Q||M)$$

where $M = \frac{1}{2}(P + Q)$. The square root of D_{JS} is a distance metric (JSD).

In our context, the JSD is useful to analyse whether a distribution over consumer segments to which item is recommended, matches the ideal distribution of consumer segments for the item, or how far the item is from the ideal distribution. This requires a notion of ideal distribution. It can be built upon past user-item interactions. It can be assumed that if enough interactions are recorded for an item, we obtain an accurate estimate of the ideal distribution, and the item has a developed profile of consumer segments. Then, if recommendations are compared to such a profile, a large JSD value indicate that these recommendations were made differently to the ideal item customer profile.

4 EVALUATION

4.1 Dataset

We analyse item exposure on the MovieLens 20M dataset [11] which contains user ratings on movies. The dataset consists of 20M ratings on 5-star scale given by 138K users on 27K movies. All users have rated at least 20 movies. No demographic information is included. As we do not evaluate the recommendation performance against a hold-out sample, the full dataset is used in our analysis.

4.2 User Segmentation

To measure item exposure, a partition into consumer segments is required. Geographic, demographic or behavioural information is commonly used by marketeers to segment consumers. As the dataset consists only of interactions between users and items, the only source for segmentation is behavioural information inferred from the interactions.

In [19], consumers are partitioned into segments picked by the X-means clustering algorithm, which is an extension of the k-means algorithm. Clustering is run on user feature vectors coming from a matrix factorisation. Each of the resulting segments represents a common taste. This approach assumes that tastes are exclusive and users belong only to one segment. Other disadvantage of this approach is its hard reproducibility as X-means clustering is highly dependent on initialisation conditions.

A more natural assumption is that users have multiple interests, with different preferences towards them. Considering the movie domain, a user might be interested in e.g. comedy and drama movies, but may prefer dramas more. Such interests might be more complex and abstract than just genres, and we think more generally of user interests as abstractions of general user behaviours. In [12], Hofmann proposed a statistical modelling technique that decomposes user preferences through latent class variables. This method has been successfully used in the recommendation task to predict the likelihood of an item being chosen by a user, according to the aspect model:

$$p(i|u) = \sum_{a \in \mathcal{A}} p(i|a)p(a|u)$$

where $\mathcal{A}$ is a set of latent aspects representing tastes/user communities, $p(i|a)$ is the probability of picking item i if aspect a has been chosen, and $p(a|u)$ holds the user preference towards the aspect. Hofmann showed that the above model can also be used for mining overlapping user communities, where aspect a represents a community and $p(a|u)$ represents a user's association to the community. For each item i, its exposure over consumer segments can be expressed as the exposure of the user group to whom it is relevant, $\mathcal{U}_i$:

$$P(c) = p(a|u \in \mathcal{U}_i) = \sum_{u \in \mathcal{U}_i} p(a|u)p(u|\mathcal{U}_i)$$

where $p(u|\mathcal{U}_i)$ is the relative frequency of obtaining interactions from user u. This can be proportional to the size of user's profile, or uniform, $\frac{1}{|\mathcal{U}_i|}$, which we adopt here. Taking this distribution as the item's ideal audience, we can compare it against the audience targeted by the recommender algorithm, which can be computed by using $\mathcal{R}_i$ instead of $\mathcal{U}_i$ above.

4.3 Experiment Setup

We analyse item exposure of recommendations produced by three well-known collaborative filtering algorithms: item- (IB; 10 neighbours) and user-based (UB; 100 neighbours) neighbourhood methods [9] and matrix factorisation (MF; 50 factors) [13]. We simulate the process of consuming recommendations by recommending 10 items to each user from which a random one is selected. Selected item is added to the training set and the whole process it repeated 100 times. Selected items form the $\mathcal{R}_i$ sets which are used to analyse the items' exposure in recommendations. The process is repeated with EPC re-ranker applied (as described in Sec. 2) to see if improving sales diversity improves performance in terms of item exposure over consumer segments. Re-ranker is set up with $\lambda = 0.5$ to give equal importance to relevance and novelty. The RankSys framework (http://ranksys.org) has been used to generate recommendations, and train the aspect model used in the user segmentation.

To mitigate the problem of measuring exposure of items without developed profiles over consumer segments, or without enough recommendations to observe exposure, we ignore items that have been consumed less than 100 times. Also, when comparing exposures produced by original and re-ranked recommendations, in order to perform paired statistical test, we focus on common items.

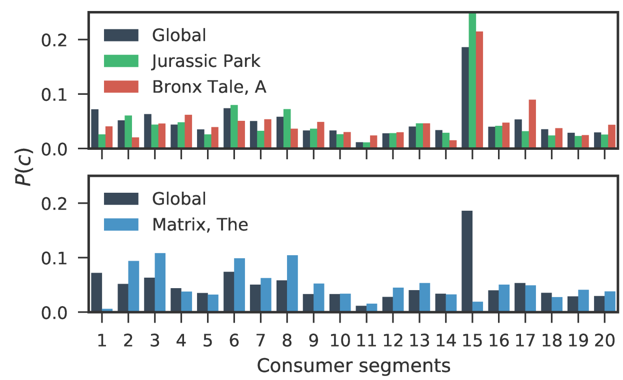


Figure 1: Exposure over consumer segments. Global exposure is compared with sample movies, showing similar (top) exposure to the global, and different (bottom).

4.4 Results

In Fig. 1 global exposure over consumer segments is plotted, which can be seen as a hypothetical item that is exposed to all users of the system. An item that is randomly exposed to users, without a specific audience, would follow the global distribution. This is confirmed in Fig. 2 where items exposed by a random recommender follow the global, which is reflected in very low JSD values. It can also be seen that items recorded in the dataset are far from the random, suggesting that they have their own, specialised audiences.

The exposure of three sample items, *Jurassic Park*, *A Bronx Tale* and *The Matrix*, is shown in Fig. 1. First two are in the top closest items to the global, with JSD of 0.15 and 0.3 for the third one. If we compare all three items, we see that they all do not match perfectly the global, however first two follow it to some extend. For *The Matrix* movie, more significant differences can be observed – two major segments, 1 and 15, have been diminished, and exposure moved to other segments, e.g. 2, 3, 6 and 8.

A correlation between item popularity and JS distance to global has been observed: more popular items are generally closer to the global. This is expected if whilst increasing popularity items are exposed to users at random. However there are plenty of interesting examples where this is not the case, e.g. *A Bronx Tale* has similar characteristic as popular *Jurassic Park* while being 10 times less popular (exposed to only 3% of users), and *The Matrix* being exposed to 40% and still having a specific audience, different than the global.

If recommendation systems were unaware of items' audiences, exposure would be following the global distribution of users across segments. Distribution over items' JSD values showed in Fig. 3a

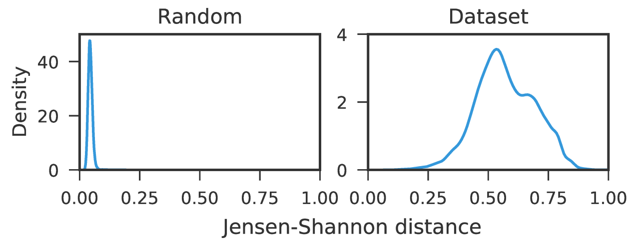
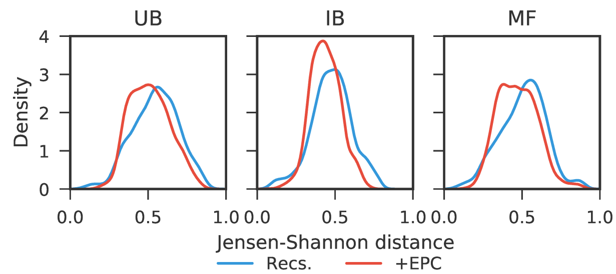
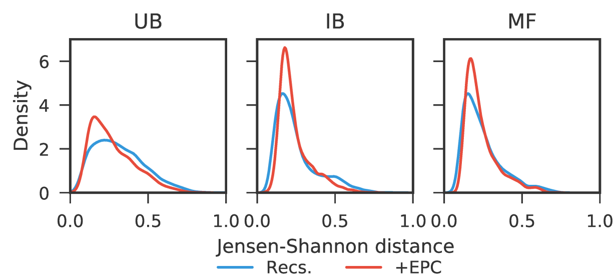


Figure 2: Distribution of JSD for random recommendations and the dataset, with the global distribution taken as the reference.



(a) Reference exposure: global.



(b) Reference exposure: dataset.

Figure 3: Distribution of JSD for recommendations coming from UB, IB and MF, using different reference exposures.

suggests that state-of-the-art recommenders (blue lines) expose differently then the global exposure, thus they learn non-random item preferences towards consumer segments. Averages of JSDs for UB, IB and MF are in range of 0.5-0.55 which shows a strong disagreements with the global.

Knowing that items have specific audiences and that recommender systems are not exposing items to users at random, we wonder if recommendations follow exposures seen in past interactions. From Fig. 3b we can read that recommendations are closer to exposures built upon past interactions than the global, however they still do not follow them perfectly and for many items the exposure is greatly different – averages of JSDs of approx. 0.25-0.3.

If we compare recommendations against each other we can observe that that exposures coming from IB and MF are concentrated closer to 0 than those from UB. We suspect that this comes from the way how the UB works – UB finds best items based on target user's neighbours which most likely share consumer segment associations. If an item is recommended to the similar neighbourhoods, this might interfere the expected exposure of that item by overexposing some consumer segments and making the exposure distant from the expected one.

As items' exposures do not ideally follow past seen in the dataset, we wonder if promoting less popular items (using the EPC greedy re-ranker) would improve not only utilisation of the catalogue but also exposure over consumer segments. Fig. 3a shows that EPC moved items towards the global exposure, which suggests adding random exposure. Wilcoxon signed-rank test of log differences in JSD showed statistically significant changes, with $\alpha = 0.05$. The average improvement is approx. 0.04.

Similarly, we compared exposures against the past interactions – small changes can be seen in Fig. 3b, with largest on the UB recommendations. For the UB, re-rankers moved the recommendations closer to the dataset exposure significantly (according to the statistical test) by 0.05 on average. The average change for IB and MF is approx. 0.01 and not statistically significant. It is interesting that re-rankers only change exposure for recommendations coming from the UB. As we suspect items in UB to be overexposed to some segments, EPC re-ranker by introducing randomness might be reducing the bias effect and making the exposures a bit closer to those based on past interactions. However, even if exposures of all items were reduced by 0.05, they still would be greatly different than the expected ones, suggesting that existing methods do not improve exposure over consumer segments.

5 CONCLUSIONS

In this paper we proposed to evaluate the exposure of items to different user types, instead of measuring just items coverage. Such analysis helps us understand how each user type perceive an item, and if recommender systems expose items to correct users. Our empirical analysis of the MovieLens dataset and state-of-the-art recommendation techniques showed that movies indeed have specific audiences however these are not targeted and reached by the recommendations. We also showed that sales diversity optimisation does not help much in reaching the expected audiences. Further work could focus on methods of improving item exposure over consumer segments in recommendations.

ACKNOWLEDGMENTS

This project has been funded by Science Foundation Ireland under Grant No. SFI/12/RC/2289.

REFERENCES

[1] P. Adamopoulos and A. Tuzhilin. 2014. On Over-specialization and Concentration Bias of Recommendations: Probabilistic Neighborhood Selection in Collaborative Filtering Systems. In *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys '14)*. ACM, New York, NY, USA, 153–160. https://doi.org/10.1145/2645710.2645752

[2] P. Adamopoulos, A. Tuzhilin, and P. Mountanos. 2015. Measuring the concentration reinforcement bias of recommender systems. *CEUR Workshop Proc.* 1441 (2015).

[3] G. Adomavicius and Y. Kwon. 2011. *Maximizing aggregate recommendation diversity: A graph-theoretic approach.* Vol. 816. CEUR-WS, 3–10.

[4] G. Adomavicius and Y. Kwon. 2012. Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques. *IEEE Trans. on Knowl. and Data Eng.* 24, 5 (May 2012), 896–911. https://doi.org/10.1109/TKDE.2011.15

[5] C. Anderson. 2006. *The Long Tail: Why the Future of Business Is Selling Less of More.* Hyperion.

[6] E. Brynjolfsson, Y. U. Hu, and M. D. Smith. 2006. From Niches to Riches: Anatomy of the Long Tail. *MIT SLOAN MANAGEMENT REVIEW* 47, 4 (2006), 67–71.

[7] O. Celma. 2010. *The Long Tail in Recommender Systems.* Springer Berlin Heidelberg, Berlin, Heidelberg, 87–107. https://doi.org/10.1007/978-3-642-13287-2_4

[8] O. Celma and P. Herrera. 2008. A New Approach to Evaluating Novel Recommendations. In *Proceedings of the 2008 ACM Conference on Recommender Systems (RecSys '08)*. ACM, New York, NY, USA, 179–186. https://doi.org/10.1145/1454008.1454038

[9] C. Desrosiers and G. Karypis. 2011. *A Comprehensive Survey of Neighborhood-based Recommendation Methods.* Springer US, Boston, MA, 107–144. https://doi.org/10.1007/978-0-387-85820-3_4

[10] D. Fleder and K. Hosanagar. 2009. Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity. *Manage. Sci.* 55, 5 (May 2009), 697–712. https://doi.org/10.1287/mnsc.1080.0974

[11] F. M. Harper and J. A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4, Article 19 (Dec. 2015), 19 pages. https://doi.org/10.1145/2827872

[12] T. Hofmann. 2004. Latent semantic models for collaborative filtering. *ACM Trans. Inf. Syst. (ACM TOIS)* 22, 1 (2004), 89–115. https://doi.org/10.1145/963770.963774

[13] Y. Hu, Y. Koren, and C. Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. (2008). https://doi.org/10.1109/ICDM.2008.22

[14] K. Niemann and M. Wolpers. 2013. A New Collaborative Filtering Approach for Increasing the Aggregate Diversity of Recommender Systems. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '13)*. ACM, New York, NY, USA, 955–963. https://doi.org/10.1145/2487575.2487656

[15] G. Shani and A. Gunawardana. 2011. *Evaluating Recommendation Systems.* Springer US, Boston, MA, 257–297. https://doi.org/10.1007/978-0-387-85820-3_8

[16] Z. Szlavik, W. J. Kowalczyk, and M. C. Schut. 2011. *Diversity measurement of recommender systems under different user choice models.*

[17] S. Vargas and P. Castells. 2011. Rank and relevance in novelty and diversity metrics for recommender systems. *Proc. fifth ACM Conf. Recomm. Syst. - RecSys '11* (2011), 109. https://doi.org/10.1145/2043932.2043955

[18] S. Vargas and P. Castells. 2014. Improving Sales Diversity by Recommending Users to Items. In *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys '14)*. ACM, New York, NY, USA, 145–152. https://doi.org/10.1145/2645710.2645744

[19] J. Wasilewski and N. Hurley. 2017. How Diverse Is Your Audience? Exploring Consumer Diversity in Recommender Systems. *RecSys 2017 Poster Proceedings* (2017).

[20] T. Zhou, Z. Kuscsik, J. Liu, M. Medo, J. R. Wakeling, and Y. Zhang. 2010. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences* 107, 10 (2010), 4511–4515. https://doi.org/10.1073/pnas.1000488107