Simulations for novel problems in recommendation: analyzing misinformation and data characteristics

ALEJANDRO BELLOGÍN[*], Universidad Autónoma de Madrid, Spain
YASHAR DELDJOO[*], Politecnico di Bari, Italy

In this position paper, we discuss recent applications of simulation approaches for recommender systems tasks. In particular, we describe how they were used to analyze the problem of misinformation spreading and understand which data characteristics affect the performance of recommendation algorithms more significantly. We also present potential lines of future work where simulation methods could advance the work in the recommendation community.

CCS Concepts: • **Information systems** → **Recommender systems**; *Collaborative filtering*.

Additional Key Words and Phrases: evaluation, data characteristics, misinformation, preference sampling

ACM Reference Format:
Alejandro Bellogín and Yashar Deldjoo. 2021. Simulations for novel problems in recommendation: analyzing misinformation and data characteristics. In *SimuRec '21: The SimuRec workshop held in conjunction with the 15th ACM Conference on Recommender Systems (RecSys), Sep 27– Oct 01, 2021, Amsterdam, Netherlands (online).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/1122445.1122456

1 IMPORTANCE OF SAMPLING IN RECOMMENDATION

Recent years have witnessed an explosion where simulations have been used in several aspects of Recommender Systems (RS), either inspired from Machine Learning (ML) or Information Retrieval (IR) problems – where simulations have been used for a long time, as in click models [10] or learning to rank [2] approaches –, or to attack inherent and concrete issues that are prevalent in RS, such as data scarcity and sparsity.

In general, sampling in recommendation has been used to open up (or *simulate*) evaluation scenarios. For example, in [22] the authors used it to downsample the observed interactions and generate different scenarios of cold-start. We also found works where sampling is employed to analyze and understand the evaluation procedure, as in [27] where it is used to perform a robustness analysis, in [3] where the authors extend the previous work to study hyperparameter optimization, or to directly debias the evaluation, as described in [8].

However, there is also a growing body of literature that has been using sampling for learning preferences, as in the well-known Bayesian Personalized Ranking (BPR) algorithm [25], which uses sampling to select pairs of items which are then provided to the algorithm. However, it is not obvious how this sampling should be done so that it always work, and several approaches have been investigated which differ in effectiveness and biases – such as popularity – learned by the method [9, 21]. Recent works that also rely on samplings or simulations are those where reinforcement learning is applied to recommendation [18] or to its evaluation [20]. Similarly, whenever the IR or ML approaches mentioned

[*] Both authors contributed equally to this research.

Presented at the SimuRec workshop held in conjunction with the 15th ACM Conference on Recommender Systems (RecSys), 2021, in Amsterdam, Netherlands.

Algorithm 1: Ratio-based user profile generator

```
Function generate user u, ratio r
    neg ⟵ { i ∈ u: i  is misinformative  } ;          // Negative claims
    neu ⟵ u \ neg ;                                   // Neutral claims
    desNeg ⟵ r · |u| ;                                // Desired negative ratio
    desNeu ⟵ (1 - r) · |u| ;                          // Desired neutral ratio
    while  (desNeg > |neg|) OR (desNeu > |neu|) do
        if  desNeg > |neg| ;                          // Downsampling negative
         then
            desNeg ⟵ desNeg - 1;
        end
        if  desNeu > |neu| ;                          // Downsampling neutral
         then
            desNeu ⟵ desNeu - 1;
        end
        newTotal ⟵ desNeg + desNeu;
        desNeg ⟵ r · newTotal;
        desNeu ⟵ (1 - r) · newTotal;
    end
    userProfile(u) ⟵ sample(neg, desNeg) ∪ sample(neu, desNeu);
end
```

before have even been applied to recommendation, as in [19], the sampling process is key for a successful translation of the approach from one area to the other.

In the rest of this paper, we present two use cases where the authors have recently applied data sampling on recommendation tasks with different goals in each case. We later discuss the main lessons learned in this process and the advantages of using simulations or sampling strategies. We end the paper with several open questions or ideas to explore in the future.

2 SUCCESSFUL APPLICATIONS OF SAMPLING

In this section, we describe two applications of data sampling on recent works performed by the authors. Section 2.1 shows how sampling allowed to analyze the effect of misinformation in recommender systems, whereas Section 2.2 describes a sampling strategy that is used to infer the impact of data characteristics in recommendation performance.

2.1 Analyzing misinformation

In [17], we analyzed the effect that some recommendation algorithms cause on the amplification of misinformation. For this, we first created a dataset by merging information from fact-checkers and Twitter. Then, to simulate a wide range of situations – not only the one captured at the moment of collecting the data – we used Algorithm 1 to generate user profiles with a pre-defined set of constraints, in particular, the proportion of users who shared some misinformative items. In this way, we would simulate a population where 20% (or 50% or 80%) of users share this type of item. This also gives us more control on the amount of information received by the algorithms in terms of sparsity or other important dataset statistics.

Algorithm 2: Sample generation procedure

```
Function data-sampling URM user-item rating matrix
    $n_u \leftarrow$ number of users of the URM;
    $n_i \leftarrow$ number of users of the URM;
    $n_r \leftarrow$ number of ratings of the URM;
    $\tau_u \leftarrow$ constraint on average number of ratings for users;
    $\tau_i \leftarrow$ constraint on maximum number of items;
    $n \leftarrow 1$ ;
    while $n \leq N$ do
        Random shuffle the rows of the URM;
        $n_u \leftarrow rnd([100, n_u])$;
        $n_i \leftarrow rnd([100, n_i])$;
        $urm_n \leftarrow$ Selection of $n_u$, $n_i$ from shuffled URM;
        if $\frac{n_r}{n_u} < \tau_u$ or $n_i > \tau_i$ then
            $n \leftarrow n + 1$;
        end
    end
    Output: $N$ sub-datasets ($urm_n$);
end
```

It is interesting to note that simulation has been very recently proposed in [23] also to study the societal impact of recommender systems. By a general event-driven simulation model, the authors analyze some case studies and discuss the implications for reproducibility such a framework may have.

2.2 Understanding data characteristics

In [13], and as an extension of [14], we aim to better understand the impact of an array of characteristics in a dataset with respect to accuracy and fairness (in [13]) or robustness (in [14]). With this goal in mind, we develop an explanatory framework using regression models in its core, a methodology originally proposed in [1]. In this framework, we test whether a set of data characteristics (independent variables of the regression model) are related to a given performance metric (the dependent variable). To obtain enough points so that the framework could produce significant results, we need to *simulate* N different datasets ($N = 600$ in both our papers). These simulated datasets are generated from sampling the original dataset as described in Algorithm2, which produces smaller datasets with slightly different characteristics in each case, although satisfying some constraints that allow the dataset to be useful in the analysis – for example, by restricting the maximum number of items (via τ_i) on these smaller datasets.

3 DISCUSSION

From the applications presented in Section 2 we have learned that simulation, and in particular, data sampling, is complex, even though it might be beneficial – sometimes the only tool to produce the data points needed for an analysis. Its difficulty relies on sampling the data in a significant but fair and realistic way, without introducing new biases, as acknowledged in recent works such as [21]. For example, in the case described in Section2.2 we experimented with imposing additional constraints to obtain simulated datasets with a particular density, number of ratings per user, or number of items.

Based on our experience, we foresee a continuous use of simulation techniques, such as data sampling or synthetic data, to experiment with novel evaluation conditions, such as those described before. In particular, consider the limitations of public datasets, with a fixed number of attributes for users and items and a (sometimes) a small number of interactions, these strategies will allow evaluating new conditions with little effort – or at least, with less effort than that of creating datasets including all the required information.

4 OPEN QUESTIONS

Throughout this position paper, we have presented several situations where simulations or sampling strategies have been successfully applied to recommendation. However, we believe there is room for improvement, and several aspects remain unexplored, both in our works and in the community, under a more general perspective. First, regarding how to extend our sampling approaches, it would be more realistic if the temporal dimension is incorporated in the process, so that the interactions follow a *compatible* order with the one in the original dataset. When done properly, this would allow generating and studying feedback loops, like those created by reinforcement learning algorithms, but at a higher, more global level. As observed recently, and in agreement with our misinformation analysis [17], some algorithms are more prone to reproduce biases at each feedback loop [24].

Regarding our second application (see Section 2.2), adding more variability and flexibility in the types of datasets being generated would allow us to address more complex questions. In particular, we envision a definition of user types (or *personas*) which are then simulated or sampled at different rates, either randomly or controlled via some parameter. Similarly, generating content features realistically would help go beyond collaborative filtering algorithms and test content-based methods at varying levels of information, quality, and sparsity [15]. For this, recent advances from the Natural Language Processing community could be convenient, where Neural Networks may generate realistic pieces of text in several domains [7, 16]. Besides content attributes, including sensitive attributes in the set of controled (or simulated) information to be generated will allow to explore fairness [12] analyses in a more comprehensive way than what is being done right now. Finally, an interesting perspective that could be promising is to shift the focus of the application, as mentioned earlier, from the data to the algorithms so that the sampling strategy instead of sampling data should sample algorithms. In this way, the input data would be fixed, and the data points for the regression analysis would be obtained from a wide selection of algorithms, probably with different hyperparameters, to increase their variability. Such an approach would have connections with automatic hyperparameter tuning techniques like Bayes Optimization [6], but it would be applied to solve a different problem and in a more extensive search space, as the type of algorithm will also be part of the sampled variables [5].

From a more general perspective, we believe several open questions need to be considered when doing simulations or, in particular, some data sampling. An important aspect that should be considered is that of reproducibility [11]. While usually sampling – or simulations in general – are random in nature, this hinders reproducing other people's works. However, by sharing the code – where customizable seeds are included wherever is needed – and/or the generated simulated/sampled datasets or the scripts used, the potential to reproduce these types of works should increase [4]. By addressing the reproducibility problem, a related issue we have also striven to present properly in the past could be solved. Here we refer to present all the technical decisions involved in a sampling strategy carefully and as detailed as possible. Sometimes pseudocodes or algorithms do not have the granularity level needed to specify implementation details that may greatly impact how the sampling or the simulation is generated.

At the same time, although there are some works already addressing this issue (see [20]), we believe it is very important to understand the potential biases that a particular simulation may be introducing in the generated data. This may be intentional but, usually, there should be some mechanism to avoid them. Related to this problem, there should exist a definition of when a simulation can be considered faithful to the original (or intended) data, which we have referred throughout this work as *realistic*. Without such definition, we might end up with data samples that are too different (or different in specific, key aspects) from the actual data, hence not satisfying their underlying assumptions or constraints. This also applies when generating synthetic data, a task not so popular in recommendation because of its difficulty, but where some efforts have been reported in the last years [26].

REFERENCES

[1] Gediminas Adomavicius and Jingjing Zhang. 2012. Impact of data characteristics on recommender systems performance. *ACM Trans. Manag. Inf. Syst.* 3, 1 (2012), 3:1–3:17. https://doi.org/10.1145/2151163.2151166

[2] Qingyao Ai, Tao Yang, Huazheng Wang, and Jiaxin Mao. 2021. Unbiased Learning to Rank: Online or Offline? *ACM Trans. Inf. Syst.* 39, 2 (2021), 21:1–21:29. https://doi.org/10.1145/3439861

[3] Vito Walter Anelli, Tommaso Di Noia, Eugenio Di Sciascio, Claudio Pomo, and Azzurra Ragone. 2019. On the discriminative power of hyper-parameters in cross-validation and how to choose them. In *Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019, Copenhagen, Denmark, September 16-20, 2019*, Toine Bogers, Alan Said, Peter Brusilovsky, and Domonkos Tikk (Eds.). ACM, 447–451. https://doi.org/10.1145/3298689.3347010

[4] Alejandro Bellogín and Alan Said. 2021. Improving Accountability in Recommender Systems Research Through Reproducibility. *CoRR* abs/2102.00482 (2021). arXiv:2102.00482 https://arxiv.org/abs/2102.00482

[5] Yoshua Bengio, Andrea Lodi, and Antoine Prouvost. 2021. Machine learning for combinatorial optimization: A methodological tour d'horizon. *Eur. J. Oper. Res.* 290, 2 (2021), 405–421. https://doi.org/10.1016/j.ejor.2020.07.063

[6] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for Hyper-Parameter Optimization. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger (Eds.). 2546–2554. https://proceedings.neurips.cc/paper/2011/hash/86e8f7ab32cfd12577bc2619bc635690-Abstract.html

[7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[8] Diego Carraro and Derek Bridge. 2020. Debiased offline evaluation of recommender systems: a weighted-sampling approach. In *SAC '20: The 35th ACM/SIGAPP Symposium on Applied Computing, online event, [Brno, Czech Republic], March 30 - April 3, 2020*, Chih-Cheng Hung, Tomás Cerný, Dongwan Shin, and Alessio Bechini (Eds.). ACM, 1435–1442. https://doi.org/10.1145/3341105.3375759

[9] Jiawei Chen, Chengquan Jiang, Can Wang, Sheng Zhou, Yan Feng, Chun Chen, Martin Ester, and Xiangnan He. 2021. CoSam: An Efficient Collaborative Adaptive Sampler for Recommendation. *ACM Trans. Inf. Syst.* 39, 3, Article 34 (May 2021), 24 pages. https://doi.org/10.1145/3450289

[10] Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. 2015. *Click Models for Web Search*. Morgan & Claypool Publishers. https://doi.org/10.2200/S00654ED1V01Y201507ICR043

[11] Maurizio Ferrari Dacrema, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. 2021. A Troubling Analysis of Reproducibility and Progress in Recommender Systems Research. *ACM Trans. Inf. Syst.* 39, 2 (2021), 20:1–20:49. https://doi.org/10.1145/3434185

[12] Yashar Deldjoo, Vito Walter Anelli, Hamed Zamani, Alejandro Bellogin, and Tommaso Di Noia. 2021. A flexible framework for evaluating user and item fairness in recommender systems. *User Modeling and User-Adapted Interaction* (2021), 1–47.

[13] Yashar Deldjoo, Alejandro Bellogín, and Tommaso Di Noia. 2021. Explaining Recommender Systems Fairness and Accuracy through the Lens of Data Characteristics. *Inf. Process. Manag.* 58, 5 (2021). https://doi.org/10.1016/j.ipm.2021.102662

[14] Yashar Deldjoo, Tommaso Di Noia, Eugenio Di Sciascio, and Felice Antonio Merra. 2020. How Dataset Characteristics Affect the Robustness of Collaborative Recommendation Models. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, Jimmy Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 951–960. https://doi.org/10.1145/3397271.3401046

[15] Yashar Deldjoo, Markus Schedl, and Peter Knees. 2021. Content-driven Music Recommendation: Evolution, State of the Art, and Challenges. *arXiv preprint arXiv:2107.11803* (2021).

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[17] Miriam Fernández, Alejandro Bellogín, and Iván Cantador. 2021. Analysing the Effect of Recommendation Algorithms on the Amplification of Misinformation. *CoRR* abs/2103.14748 (2021). arXiv:2103.14748 https://arxiv.org/abs/2103.14748

[18] Dorota Glowacka. 2019. Bandit Algorithms in Information Retrieval. *Found. Trends Inf. Retr.* 13, 4 (2019), 299–424. https://doi.org/10.1561/1500000067

[19] Katja Hofmann, Anne Schuth, Alejandro Bellogín, and Maarten de Rijke. 2014. Effects of Position Bias on Click-Based Recommender Evaluation. In *Advances in Information Retrieval - 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings (Lecture Notes in Computer Science, Vol. 8416)*, Maarten de Rijke, Tom Kenter, Arjen P. de Vries, ChengXiang Zhai, Franciska de Jong, Kira Radinsky, and Katja Hofmann (Eds.). Springer, 624–630. https://doi.org/10.1007/978-3-319-06028-6_67

[20] Jin Huang, Harrie Oosterhuis, Maarten de Rijke, and Herke van Hoof. 2020. Keeping Dataset Biases out of the Simulation: A Debiased Simulator for Reinforcement Learning based Recommender Systems. In *RecSys 2020: Fourteenth ACM Conference on Recommender Systems, Virtual Event, Brazil, September 22-26, 2020*, Rodrygo L. T. Santos, Leandro Balby Marinho, Elizabeth M. Daly, Li Chen, Kim Falk, Noam Koenigstein, and Edleno Silva de Moura (Eds.). ACM, 190–199. https://doi.org/10.1145/3383313.3412252

[21] Amir Jadidinejad, Craig Macdonald, and Iadh Ounis. 2019. How Sensitive is Recommendation Systems' Offline Evaluation to Popularity?. In *Proceedings of the Workshop on Offline Evaluation for Recommender Systems (REVEAL '19), co-located with the 13th ACM Conference on Recommender Systems (RecSys 2019)*.

[22] Daniel Kluver and Joseph A. Konstan. 2014. Evaluating recommender behavior for new users. In *Eighth ACM Conference on Recommender Systems, RecSys '14, Foster City, Silicon Valley, CA, USA - October 06 - 10, 2014*, Alfred Kobsa, Michelle X. Zhou, Martin Ester, and Yehuda Koren (Eds.). ACM, 121–128. https://doi.org/10.1145/2645710.2645742

[23] Eli Lucherini, Matthew Sun, Amy Winecoff, and Arvind Narayanan. 2021. T-RECS: A Simulation Tool to Study the Societal Impact of Recommender Systems. *CoRR* abs/2107.08959 (2021). arXiv:2107.08959 https://arxiv.org/abs/2107.08959

[24] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. Feedback Loop and Bias Amplification in Recommender Systems. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux (Eds.). ACM, 2145–2148. https://doi.org/10.1145/3340531.3412152

[25] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009*, Jeff A. Bilmes and Andrew Y. Ng (Eds.). AUAI Press, 452–461. https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=1630&proceeding_id=25

[26] Manel Slokom. 2018. Comparing recommender systems using synthetic data. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, Sole Pera, Michael D. Ekstrand, Xavier Amatriain, and John O'Donovan (Eds.). ACM, 548–552. https://doi.org/10.1145/3240323.3240325

[27] Daniel Valcarce, Alejandro Bellogín, Javier Parapar, and Pablo Castells. 2018. On the robustness and discriminative power of information retrieval metrics for top-N recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, Sole Pera, Michael D. Ekstrand, Xavier Amatriain, and John O'Donovan (Eds.). ACM, 260–268. https://doi.org/10.1145/3240323.3240347