

# Graph-based Explainable Recommendation Systems: Are We Rigorously Evaluating Explanations?

A Position Paper.

Andrea Montagna<sup>1,2,\*</sup>, Alvisè De Biasio<sup>1,2,†</sup>, Nicolò Navarin<sup>1</sup> and Fabio Aiolli<sup>1</sup>

<sup>1</sup>Department of Mathematics, University of Padova, Via Trieste 63, Padova, 35131 PD, Italy

<sup>2</sup>R&D Department, estilos srl, Via Ca' Marcello 67/D, Venezia, 30172 VE, Italy

## Abstract

In recent years, we have witnessed an increase in the amount of published research in the field of Explainable Recommender Systems. These systems are designed to help users find the items of the most interest by providing not only suggestions but also the reasons behind those recommendations. Research has shown that there are many advantages to complementing a recommendation with a convincing explanation. For example, such an explanation can often lead to an increase in user trust, which in turn can improve recommendation effectiveness and system adoption. In particular, for this reason, many research works are studying explainable recommendation algorithms based on graphs, e.g., exploiting knowledge graphs or graph neural networks based methods. The use of graphs is very promising since algorithms can, in principle, combine the benefits of personalization and graph reasoning, thus potentially improving the effectiveness of both recommendations and explanations. However, although graph-based algorithms have been repeatedly shown to bring improvements in terms of ranking quality, not much literature has yet studied how to properly evaluate the quality of the corresponding explanations. In this position paper, we focus on this problem, examining in detail how the explanations of explainable recommenders based on graphs are currently evaluated and discussing how they could be evaluated in the future in a more quantitative and comparable way in compliance with the well-known Explainable Recommender Systems guidelines.

## Keywords

Explainability, Graph Neural Networks, Recommender Systems

## 1. Introduction

Graph-based algorithms have attracted the interest of many researchers because of the capabilities they offer to represent the world of interactions, particularly those related to humans. They are a promising field because the learning process can be based directly on graphs that, in addition to representing user-object interactions, can include contextual information such as user demographics, product categories, and other attributes. In particular with the objective of capturing these connections and exploiting these potentials through user suggestions, Knowledge

---


HCAI4U 2023: Workshop on User Perspectives in Human-Centred Artificial Intelligence, September 20, 2023, Turin, Italy.


\*Corresponding author.

†These authors contributed equally.

✉ andrea.montagna@phd.unipd.it (A. Montagna); alvisè.debiasio@phd.unipd.it (A. De Biasio); nnavarin@math.unipd.it (N. Navarin); aiolli@math.unipd.it (F. Aiolli)

ORCID 0000-0002-3237-3464 (A. Montagna); 0000-0003-0528-6223 (A. De Biasio); 0000-0002-4108-1754 (N. Navarin)

 © 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Graphs (KGs) and Graph Neural Networks (GNNs) have gained significant attention in recent years, especially in the Recommender Systems (RSs) field to ensure fairness [1], improve business value [2], or generate relevant, yet explainable, recommendations to users from graphs [3, 4]. As users have increasingly demanded explanations for recommendations in recent years, it becomes crucial for model developers to provide insights into how and why those recommendations are made. Unlike traditional recommender systems, this kind of approach allows any recommendation to be generated by simultaneously integrating it with a corresponding explanation. The explanations provided can be very persuasive [5, 6, 7] in that they often exploit graph reasoning logic that allows the explanation to be represented as a path on the graph. For example, an explanation for a “Back to the Future” recommendation, may be intuitively represented with the path “user  $\xrightarrow{\text{watched}}$  Forrest Gump  $\xrightarrow{\text{directed by}}$  Robert Zemeckis  $\xrightarrow{\text{that also directed}}$  Back to the Future” [8]. However, although graph-based methods have repeatedly shown remarkable performance in modeling complex user-item relational dependencies and generating relevant recommendations, the transparency and interpretability of the underlying reasoning process still remain a significant concern. In particular, the main problem lies in establishing a clear understanding of what a Graph-Based Explainable Recommender System (GxRS) should provide as an explanation, how it could be useful for the final users, and especially how accurate the outcome is.

In the following sections, we will examine several works reported in the most recent surveys [9, 10, 11, 12, 13, 14] in the field of GNN- and KG-based recommender systems, from 2018 to 2023 (see Table 1). What emerges from our analysis is that many existing studies have focused on the use of graphs to improve the performance of recommender systems in terms of recommendation quality, diversity, and other conventional measures, while less rigorously assessing the quality of the corresponding explanations. Studies often offer only qualitative case-based examples where a particular explanation is represented graphically as a path on the graph. However, what is generally lacking is a quantitative and comparable evaluation of the quality of explanations according to the widely known guidelines of explainable recommendation systems [15]. While graphical explanations are useful for getting an intuitive idea of the underlying reasoning process, they also severely limit the comparison of algorithms and, more generally, the progress in this particular field. Instead, it may be beneficial for research to focus more in the future on the evaluation aspects by designing metrics to provide quantitative insights into the complete decision-making process to ensure that the algorithms’ explanations are useful in practice from a human perspective.

To clarify the challenges mentioned above, the remainder of the article is organized as follows. In Section 2 we offer a general introduction to recent graph-based explainable recommendation algorithms. Subsequently, in Section 3 we focus on the current methodologies used in the literature to evaluate these algorithms. Then, in Section 4 we bring to the community’s attention the importance of the well-known explainable recommender systems guidelines for evaluation, while also discussing potential ways to use these guidelines in future research works. Finally, Section 5 concludes the article with a summary of the findings. We hope that our work will improve the understanding of current advances, identifying major challenges, and encouraging the development of more robust and user-oriented approaches for evaluating graph-based explainable recommender systems in the future.

**Table 1**

Research works on graph-based explainable recommender systems that emerged from the surveyed literature [9, 10, 11, 12, 13] divided by algorithmic method and explainability evaluation approach.

Algorithmic Method	No Explainability Evaluation	Qualitative Explainability Evaluation	Quantitative Explainability Evaluation
Embedding-based	[16, 17, 18, 19, 20]	[21, 22, 23]	
Path-based		[24, 25, 26, 27, 28, 29, 30, 8, 31, 32, 33, 34, 35]	[36, 37, 38, 39, 40, 41]

## 2. Graph-based Explainable Recommendation Algorithms

In the literature, a number of different studies proposed the use of GNN- or KG-based algorithms to generate relevant, yet explainable recommendations from graphs. These graphs often provide additional information in addition to the most commonly used user-item interactions, including demographic characteristics of the user (e.g., age, gender), various attributes of the item (e.g., product category, price range), and contextual features (e.g., time, location) interconnected in a graph. A graph can be classified into homogeneous (if the edge connects only two nodes and there is only one type of nodes and edges), heterogeneous (if the edge connects only two nodes and there are multiple types of nodes and edges), or hypergraph (where each edge joins more than two nodes). While edges represent a relation (interaction or property) of the node, each node represents an entity of the dataset that could have one or more associated properties and could interact with other entities. There are various techniques for using such graph-based information for recommendations and/or explanation purposes.

Depending on how the graph is handled in the learning process, we can distinguish different graph-based explainability recommendation techniques in the literature. For example, in certain cases, neural networks can be exploited to decompose the graph in the form of embeddings or paths (see Table 1). In particular, *embedding-based methods* typically aim to learn embedding representations of users, items, and other entities from the graph that can be used to generate recommendations or explanations. However, embedding-based methods generally lack the ability to discover multi-hop relational paths from the graph to generate explanations. Therefore, the explanations provided are generated by exploiting empirical criteria of similarity matching among the various embeddings in the graph to motivate post-hoc a given recommendation (*weak explainability*). Instead, *path-based methods* first identify connectivity paths between users and items and then feed those paths into the recommendation algorithms to generate recommendations and explanations. However, although the explanations provided by these models often appear quite convincing, as they are based on complex multi-hop reasoning, considering all possible paths between a given user-item pair may involve irrelevant ones that can lead to mismatches with real user preferences (*error propagation*). Given the current limitations of *pure* path-based and embedding-based algorithms, other hybrid [8, 31, 32] algorithms have also been studied in the literature. These methodologies should, in principle, improve recommendations and explanations by alleviating the weak explainability and error propagation problems. However, as we discuss in the next section, explanations are not always rigorously evaluated.

Considering the above open challenges, it is highly important to effectively evaluate not only the quality of the recommendations of current methods but also the quality of the corresponding explanations. For example, some algorithms may be more suitable for certain application domains because they may provide better explanations. Other algorithms may be preferable for other contexts because they may provide more relevant yet explainable recommendations.

### 3. Are we Really Evaluating the Quality of Explanations of Graph-based Explainable Recommendation Algorithms?

In the above literature, a variety of methods and metrics are used for evaluation purposes. In particular, all the surveyed studies employ well-known offline metrics (e.g., Precision, Recall, NDCG, AUC) from the RSs literature to evaluate the relevance of recommendations<sup>1</sup>. These metrics are typically used to evaluate the performance of a RS in recommending items of most interest to users. Proposed graph-based algorithms are often able to beat baselines in terms of relevance or other well-known quality factors such as diversity and coverage because, since graphs are often used as additional contextual information to the user-item interaction matrix, they allow, in principle, more accurate recommendations to be generated. However, especially for an Explainable Recommender Systems, while it is important to assess the relevance of recommendations, it should also be equally important to assess the quality of the corresponding explanations. Indeed, the recommendation algorithm could, in principle, produce good-quality recommendations but weak explanations.

Unfortunately, when analyzing the above literature in detail, it emerges that only a few papers [36, 37, 38, 39, 40, 41] have evaluated the quality of explanations as rigorously as they have assessed recommendations relevance. Indeed, in terms of explanation quality, almost all studies (see Table 1) provide some qualitative case-based analysis to intuitively evaluate the quality of the algorithmic reasoning process. Typically, a specific recommendation of a certain item is selected for a given user, and a *graphical representation* of the explanation provided by the algorithm is proposed. Supplementing the graphical representation, some empirical observations are often provided to state that, at least intuitively, the considered explanation seems realistic. However, what is generally missing is a quantitative and comparable assessment of the quality of the system’s overall explanations, i.e., a goal-oriented evaluation based on different factors of the explanations that the system should provide for each recommendation to every user, as is typically done instead when assessing recommendations relevance.

#### 3.1. Current Evaluation of Explanations Quality in Graph-Based Recommender Systems

Besides the typically employed qualitative case analyses, only a few articles, among the ones listed in Table 1, proposed to evaluate the explanations of the proposed Explainable Recommender Systems in a more quantitative way. For example, Lyu et al. [38] used *ROUGE* to

---

<sup>1</sup>We refer readers to some recent surveys [42, 43] for further insights on well-known offline evaluation metrics that are widely used in the RSs literature to assess the relevance of recommendations.

evaluate explanations offline. The metric is typically used for the evaluation of text summarization tasks and measures the number of overlapping words between the generated text and the ground truth. Since in the paper, the explanations generated by the algorithm are expressed in natural language, the authors can use the metric to assess how close these explanations are to the ground truth user reviews. However, the metric can only be used to evaluate the natural language-based explanation style, which is a recent area of research in the literature. Therefore, the proposed evaluation is not suitable for evaluating the more widely-adopted path-based explanation style, i.e., where the logical reasoning of the algorithm is represented as a path on the graph. To overcome this limitation, a similar methodology recently employed by Tai et al. [40] and Zhao et al. [41] consists of evaluating the ability of an algorithm to provide explanation paths that contain entities also present in the form of words in user reviews, exploiting well-known relevance-based metrics such as NDCG and Recall. However, as in the previous case, if user reviews are not present, this kind of evaluation methodology is not applicable. Exploiting a different methodology, Ma et al. [37] proposed to evaluate the quality of the explanations (in terms of relevance and diversity) from a *human* perspective. In particular, the authors randomly selected 100 user-item recommendation pairs and the corresponding path-based explanations generated by the recommender system. Then they selected 10 human raters who have machine learning experience to manually evaluate the quality of explanations. However, as is also known in other areas in the field of recommender systems, this particular online evaluation methodology can be very expensive to perform on a large scale and subject to user bias. Hence, the overall validity of the final results may be compromised if the human raters are not carefully selected. Another methodology has been proposed recently by Wang et al. [36]. In particular, given a certain explanation for a user recommendation, proposed to evaluate the degree to which the explanation path conforms to the particular user profile. Specifically, for a given user, the authors first construct the user profile containing his/her interactions with the entities of the graph. Then they measured the number of entities in the explanation path that are also present in the user profile. Moreover, since an explanation path can be based on multiple hops between different graph entities, very long reasoning paths would be able to match more user profile entities. Correspondingly, the authors' proposed evaluation is based on a hyperparameter that considers only a certain number of entities in the explanation path for evaluation purposes. However, as noted by the authors, this evaluation methodology is very inefficient. Hence, they sampled only 100 test set users and evaluated the explanations of the top-20 recommendations for each of them. Finally, in another recent work by Geng et al. [39] it has been proposed to measure through the New Reach Ratio ( $NR^2$ ) metric in which terms a graph-based explainable model is able to mitigate the recall bias.

#### **4. Towards a More Standardized Evaluation in Compliance with the Recommender Systems Explainability Guidelines**

In the previous section, we highlighted that most of the works in the GxRS literature used a qualitative case-based analysis rather than a quantitative approach to intuitively evaluate the explanations provided by the model. Unfortunately, the lack of adoption of a quantitative and comparable framework for styling, presenting, personalizing, and evaluating such explanations,

**Table 2**

Existing metrics used in the literature to evaluate the quality of explanations.

Metrics	References
MEP, MER	[44]
Model Fidelity	[45]
EP, E-NDCG	[46]
PN, PS	[47]
LIR, LID, SEP, SED, PTD, PTC	[48, 49]
$NR^2$	[39]

does not allow to compare the different models in terms of explainability results. The adoption of shared guidelines that employ quantitative metrics would allow this issue to be solved.

In 2015 Tintarev and Masthoff [15], and more recently Chen et al. [50] and Mohseni et al. [51] released well-known guidelines to create a common evaluation framework for Explainable Recommender Systems (of which GxRS are part). The guidelines provide a formal process for assessing the explainability of a model. Following this process, the developer can define the goal of the model, the target user, and the evaluation metrics to determine how much the model performs in terms of explainability, considering style, presentation, personalization, and evaluation aspects. For example, in terms of *presentation*, the recommendations provider may desire to give explanations that are structured in a certain way. Currently, most of the explanations are provided to users by using a template-based structure (e.g., indicating how many similar users have the same tastes of the current user), a graphical representation (e.g., considering a path on a graph) or natural language [52, 53]. Instead, in terms of *evaluation*, the quality of explanations is typically assessed considering certain goals, as transparency, scrutability, trust, effectiveness, persuasiveness, efficiency, and satisfaction. The ability of explanations to achieve such goals can be evaluated by calculating certain metrics, considering specific case studies (e.g., for a qualitative assessment) or through online experiments (e.g., user studies, A/B tests) [50]. Each of these method has certain advantages and disadvantages. For example, qualitative case-based analyses can be used to intuitively assess whether explanations are realistic or not. However, the evaluation may be affected by bias, and the outcome of different models is not comparable. Instead, analyses based on quantitative metrics (e.g., *Probability of Necessity*, *Mean Explainability Precision*) [47, 44] could be less intuitive, but they are easy to benchmark, comparable, and more efficient.

For example, future works may exploit existing metrics proposed in the literature to evaluate the explanations quality (e.g., *Mean Explainability Precision* [44], or *Explainability Power* [46]). A representative list of explainability metrics from the literature [50, 13, 54, 55] is presented in Table 2. In particular, researchers could focus on balancing multiple perspective at the same time, such as fairness and explainability measuring the Path and Diversity Scores proposed by Fu et al. [56]. Moreover, in the future it could be worth to investigate the relation between the recommendation list and the explanation properties as proposed by Balloccu et al.. In particular, in recent works, e.g., [48, 49], Balloccu et al. propose six novel metrics to evaluate the quality of explanations, i.e., Linking Interaction Recency (LIR), Linking Interaction Diversity(LID), Shared Entity Popularity (SEP), Shared Entity Diversity (SED), Explanation Path Type Diversity (PTD), and Explanation

Path Type Concentration (PTC). The use of such quantitative evaluation metrics can enhance the interpretability and transparency of recommendations, empowering users to make informed decisions based on comprehensible explanations. Overall the integration of such a quantitative evaluation framework may not only provide concrete evidence of model effectiveness but it may also contribute to the advancement and adoption of explainable recommender systems in real-world applications.

Besides exploiting existing metrics, other research directions may also be interesting for the future. In particular, while existing metrics can be used to measure certain aspects of explanations, future research may study other quantitative and comparable methods that can be used to evaluate the quality of explanations in compliance with the explainability guidelines, e.g., assessing explanations in terms of style, personalization and presentation aspects. Moreover, given the limitations of available datasets in reflecting real user preferences in terms of the explanations provided by the models, further research may focus on collecting datasets that have such information. Furthermore, another important aspect to consider for the future may be the inclusion of human-centered evaluation methods. Indeed, evaluating the performance and effectiveness of explainability from a human perspective is essential to gain valuable insights into the usability and impact of the explanations provided by an algorithm [57]. Finally, another effective enhancement for evaluating explanations could be providing explanations to users through an interface designed to facilitate access, increase comprehension, and collect users' feedback during the usage experience. Through this interface, providers could measure the adoption of a system that may seek to achieve one or more objectives at the same time, e.g., considering effectiveness [4], persuasiveness, scrutability aspects.

## 5. Conclusion

In this article, we discussed how the explanations provided by graph-based explainable recommendation systems are currently evaluated, pointing out open challenges and future research directions in this area concerning evaluation methods. What emerges from our analysis is that most papers evaluated the quality of explanations through a qualitative case-based analysis, while only a few articles proposed metrics for a more quantitative evaluation. Moreover, the current metrics are not sufficient to comprehensively evaluate all the different types of explanation and are only partially compliant with the well-known explainable recommender system guidelines. Future research will need to address current limitations by providing new guidelines-compliant evaluation methodologies. With this work, we encourage researchers to adopt a more quantitative and comparable approach when evaluating the quality of the explanations. We hope that our efforts will inspire further research in this field and lead to the creation of more comprehensive and guideline-compliant methods for assessing and comparing the quality of explanations of graph-based explainable recommendation algorithms.

## 6. Acknowledgments

This work was partially funded by estilos srl.

## References

- [1] E. Purificato, L. Boratto, E. W. De Luca, Do Graph Neural Networks Build Fair User Models? Assessing Disparate Impact and Mistreatment in Behavioural User Profiling, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, CIKM '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 4399–4403. URL: <https://doi.org/10.1145/3511808.3557584>. doi:10.1145/3511808.3557584.
- [2] A. De Biasio, A. Montagna, F. Aiolli, N. Navarin, A systematic review of value-aware recommender systems, *Expert Systems with Applications* 226 (2023) 120131. URL: <https://www.sciencedirect.com/science/article/pii/S0957417423006334>. doi:10.1016/j.eswa.2023.120131.
- [3] C. Scheel, A. Castellanos, T. Lee, E. W. De Luca, The Reason Why: A Survey of Explanations for Recommender Systems, in: A. Nürnberger, S. Stober, B. Larsen, M. Detyniecki (Eds.), *Adaptive Multimedia Retrieval: Semantics, Context, and Adaptation*, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2014, pp. 67–84. doi:10.1007/978-3-319-12093-5\_3.
- [4] E. Purificato, B. A. Manikandan, P. V. Karanam, M. V. Pattadkal, E. W. D. Luca, Evaluating Explainable Interfaces for a Knowledge Graph-Based Recommender System, in: P. Brusilovsky, M. d. Gemmis, A. Felfernig, E. Lex, P. Lops, G. Semeraro, M. C. Willemssen (Eds.), *Proceedings of the 8th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems*, volume 2948 of *CEUR Workshop Proceedings*, CEUR, Online Event, September, 2021, pp. 73–88. URL: <https://ceur-ws.org/Vol-2948/#paper5>, ISSN: 1613-0073.
- [5] S. Gkika, G. Lekakos, The Persuasive Role of Explanations in Recommender Systems, in: A. Öörni, S. Kelders, L. v. Gemert-Pijnen, H. Oinas-Kukkonen (Eds.), *Proceedings of the Second International Workshop on Behavior Change Support Systems*, volume 1153 of *CEUR Workshop Proceedings*, CEUR, Padua, Italy, 2014, pp. 59–68. URL: [https://ceur-ws.org/Vol-1153/#Paper\\_6](https://ceur-ws.org/Vol-1153/#Paper_6), ISSN: 1613-0073.
- [6] P. Cremonesi, F. Garzotto, R. Turrin, Investigating the Persuasion Potential of Recommender Systems from a Quality Perspective: An Empirical Study, *ACM Transactions on Interactive Intelligent Systems* 2 (2012) 11:1–11:41. URL: <https://doi.org/10.1145/2209310.2209314>. doi:10.1145/2209310.2209314.
- [7] A. De Biasio, M. Monaro, L. Oneto, L. Ballan, N. Navarin, On the problem of recommendation for sensitive users and influential items: Simultaneously maintaining interest and diversity, *Knowledge-Based Systems* 275 (2023) 110699. URL: <https://www.sciencedirect.com/science/article/pii/S0950705123004495>. doi:10.1016/j.knosys.2023.110699.
- [8] H. Wang, F. Zhang, J. Wang, M. Zhao, W. Li, X. Xie, M. Guo, RippleNet: Propagating User Preferences on the Knowledge Graph for Recommender Systems, in: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM '18*, Association for Computing Machinery, New York, NY, USA, 2018, pp. 417–426. URL: <https://dl.acm.org/doi/10.1145/3269206.3271739>. doi:10.1145/3269206.3271739.
- [9] Y. Chu, J. Yao, C. Zhou, H. Yang, Graph Neural Networks in Modern Recommender Systems, in: L. Wu, P. Cui, J. Pei, L. Zhao (Eds.), *Graph Neural Networks: Foundations*,



- Frontiers, and Applications, Springer Nature, Singapore, 2022, pp. 423–445. URL: [https://doi.org/10.1007/978-981-16-6054-2\\_19](https://doi.org/10.1007/978-981-16-6054-2_19). doi:10.1007/978-981-16-6054-2\_19.
- [10] S. Wu, F. Sun, W. Zhang, X. Xie, B. Cui, Graph Neural Networks in Recommender Systems: A Survey, *ACM Computing Surveys* 55 (2022) 97:1–97:37. URL: <https://dl.acm.org/doi/10.1145/3535101>. doi:10.1145/3535101.
- [11] C. Gao, Y. Zheng, N. Li, Y. Li, Y. Qin, J. Piao, Y. Quan, J. Chang, D. Jin, X. He, Y. Li, A Survey of Graph Neural Networks for Recommender Systems: Challenges, Methods, and Directions, *ACM Transactions on Recommender Systems* 1 (2023) 3:1–3:51. URL: <https://dl.acm.org/doi/10.1145/3568022>. doi:10.1145/3568022.
- [12] H. Yuan, H. Yu, S. Gui, S. Ji, Explainability in Graph Neural Networks: A Taxonomic Survey, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45 (2023) 5782–5799. doi:10.1109/TPAMI.2022.3204236, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.
- [13] Q. Guo, F. Zhuang, C. Qin, H. Zhu, X. Xie, H. Xiong, Q. He, A Survey on Knowledge Graph-Based Recommender Systems, *IEEE Transactions on Knowledge and Data Engineering* 34 (2022) 3549–3568. doi:10.1109/TKDE.2020.3028705, conference Name: IEEE Transactions on Knowledge and Data Engineering.
- [14] G. Balloccu, L. Boratto, C. Cancedda, G. Fenu, M. Marras, Knowledge is Power, Understanding is Impact: Utility and Beyond Goals, Explanation Quality, and Fairness in Path Reasoning Recommendation, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), *Advances in Information Retrieval, Lecture Notes in Computer Science*, Springer Nature Switzerland, Cham, 2023, pp. 3–19. doi:10.1007/978-3-031-28241-6\_1.
- [15] N. Tintarev, J. Masthoff, Explaining Recommendations: Design and Evaluation, in: F. Ricci, L. Rokach, B. Shapira (Eds.), *Recommender Systems Handbook*, Springer US, Boston, MA, 2015, pp. 353–382. URL: [https://doi.org/10.1007/978-1-4899-7637-6\\_10](https://doi.org/10.1007/978-1-4899-7637-6_10). doi:10.1007/978-1-4899-7637-6\_10.
- [16] Q. Li, X. Tang, T. Wang, H. Yang, H. Song, Unifying Task-Oriented Knowledge Graph Learning and Recommendation, *IEEE Access* 7 (2019) 115816–115828. doi:10.1109/ACCESS.2019.2932466, conference Name: IEEE Access.
- [17] X. Sha, Z. Sun, J. Zhang, Hierarchical attentive knowledge graph embedding for personalized recommendation, *Electronic Commerce Research and Applications* 48 (2021) 101071. URL: <https://www.sciencedirect.com/science/article/pii/S1567422321000430>. doi:10.1016/j.elerap.2021.101071.
- [18] C. Shi, B. Hu, W. X. Zhao, P. S. Yu, Heterogeneous Information Network Embedding for Recommendation, *IEEE Transactions on Knowledge and Data Engineering* 31 (2019) 357–370. doi:10.1109/TKDE.2018.2833443, conference Name: IEEE Transactions on Knowledge and Data Engineering.
- [19] Z. Sun, J. Yang, J. Zhang, A. Bozzon, L.-K. Huang, C. Xu, Recurrent knowledge graph embedding for effective recommendation, in: *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys '18*, Association for Computing Machinery, New York, NY, USA, 2018, pp. 297–305. URL: <https://dl.acm.org/doi/10.1145/3240323.3240361>. doi:10.1145/3240323.3240361.
- [20] X. Tang, T. Wang, H. Yang, H. Song, AKUPM: Attention-Enhanced Knowledge-Aware

User Preference Model for Recommendation, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1891–1899. URL: <https://dl.acm.org/doi/10.1145/3292500.3330705>. doi:10.1145/3292500.3330705.

- [21] Q. Ai, V. Azizi, X. Chen, Y. Zhang, Learning Heterogeneous Knowledge Base Embeddings for Explainable Recommendation, *Algorithms* 11 (2018) 137. URL: <http://www.mdpi.com/1999-4893/11/9/137>. doi:10.3390/a11090137.
- [22] Y. Cao, X. Wang, X. He, Z. Hu, T.-S. Chua, Unifying Knowledge Graph Learning and Recommendation: Towards a Better Understanding of User Preferences, in: The World Wide Web Conference, WWW '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 151–161. URL: <https://dl.acm.org/doi/10.1145/3308558.3313705>. doi:10.1145/3308558.3313705.
- [23] J. Huang, W. X. Zhao, H. Dou, J.-R. Wen, E. Y. Chang, Improving Sequential Recommendation with Knowledge-Enhanced Memory Networks, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 505–514. URL: <https://dl.acm.org/doi/10.1145/3209978.3210017>. doi:10.1145/3209978.3210017.
- [24] Z. Yang, S. Dong, HAGERec: Hierarchical Attention Graph Convolutional Network Incorporating Knowledge Graph for Explainable Recommendation, *Knowledge-Based Systems* 204 (2020) 106194. URL: <https://www.sciencedirect.com/science/article/pii/S0950705120304196>. doi:10.1016/j.knosys.2020.106194.
- [25] W. Ma, M. Zhang, Y. Cao, W. Jin, C. Wang, Y. Liu, S. Ma, X. Ren, Jointly Learning Explainable Rules for Recommendation with Knowledge Graph, in: The World Wide Web Conference, ACM, San Francisco CA USA, 2019, pp. 1210–1221. URL: <https://dl.acm.org/doi/10.1145/3308558.3313607>. doi:10.1145/3308558.3313607.
- [26] X. Wang, K. Liu, D. Wang, L. Wu, Y. Fu, X. Xie, Multi-level Recommendation Reasoning over Knowledge Graphs with Reinforcement Learning, in: Proceedings of the ACM Web Conference 2022, WWW '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 2098–2108. URL: <https://dl.acm.org/doi/10.1145/3485447.3512083>. doi:10.1145/3485447.3512083.
- [27] X. Wang, D. Wang, C. Xu, X. He, Y. Cao, T.-S. Chua, Explainable Reasoning over Knowledge Graphs for Recommendation, *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (2019) 5329–5336. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/4470>. doi:10.1609/aaai.v33i01.33015329, number: 01.
- [28] X. Huang, Q. Fang, S. Qian, J. Sang, Y. Li, C. Xu, Explainable Interaction-driven User Modeling over Knowledge Graph for Sequential Recommendation, in: Proceedings of the 27th ACM International Conference on Multimedia, MM '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 548–556. URL: <https://doi.org/10.1145/3343031.3350893>. doi:10.1145/3343031.3350893.
- [29] Y. Xian, Z. Fu, S. Muthukrishnan, G. de Melo, Y. Zhang, Reinforcement Knowledge Graph Reasoning for Explainable Recommendation, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 285–294. URL: <https://doi.org/10.1145/3331184.3331203>. doi:10.1145/3331184.3331203.

- [30] Y. Xian, Z. Fu, H. Zhao, Y. Ge, X. Chen, Q. Huang, S. Geng, Z. Qin, G. de Melo, S. Muthukrishnan, Y. Zhang, CAFE: Coarse-to-Fine Neural Symbolic Reasoning for Explainable Recommendation, *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (2020) 1645–1654. URL: <http://arxiv.org/abs/2010.15620>. doi:10.1145/3340531.3412038, 28 citations (Semantic Scholar/arXiv) [2022-02-14] 28 citations (Semantic Scholar/DOI) [2022-02-14] arXiv: 2010.15620.
- [31] X. Wang, X. He, Y. Cao, M. Liu, T.-S. Chua, KGAT: Knowledge Graph Attention Network for Recommendation, in: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19*, Association for Computing Machinery, New York, NY, USA, 2019, pp. 950–958. URL: <https://dl.acm.org/doi/10.1145/3292500.3330989>. doi:10.1145/3292500.3330989.
- [32] H. Chen, Y. Li, X. Sun, G. Xu, H. Yin, Temporal Meta-path Guided Explainable Recommendation, in: *Proceedings of the 14th ACM International Conference on Web Search and Data Mining, WSDM '21*, Association for Computing Machinery, New York, NY, USA, 2021, pp. 1056–1064. URL: <https://dl.acm.org/doi/10.1145/3437963.3441762>. doi:10.1145/3437963.3441762.
- [33] W. Song, Z. Duan, Z. Yang, H. Zhu, M. Zhang, J. Tang, Ekar: An Explainable Method for Knowledge Aware Recommendation, arXiv:1906.09506 [cs] (2022). URL: <http://arxiv.org/abs/1906.09506>, arXiv: 1906.09506.
- [34] X. Xin, X. He, Y. Zhang, Y. Zhang, J. Jose, Relational Collaborative Filtering: Modeling Multiple Item Relations for Recommendation, in: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, Paris France, 2019, pp. 125–134. URL: <https://dl.acm.org/doi/10.1145/3331184.3331188>. doi:10.1145/3331184.3331188.
- [35] B. Hu, C. Shi, W. X. Zhao, P. S. Yu, Leveraging Meta-path based Context for Top- N Recommendation with A Neural Co-Attention Model, in: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '18*, Association for Computing Machinery, New York, NY, USA, 2018, pp. 1531–1540. URL: <https://dl.acm.org/doi/10.1145/3219819.3219965>. doi:10.1145/3219819.3219965.
- [36] Q. Wang, E. Tragos, N. Hurley, B. Smyth, A. Lawlor, R. Dong, Entity-Enhanced Graph Convolutional Network for Accurate and Explainable Recommendation, in: *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization, UMAP '22*, Association for Computing Machinery, New York, NY, USA, 2022, pp. 79–88. URL: <https://dl.acm.org/doi/10.1145/3503252.3531316>. doi:10.1145/3503252.3531316.
- [37] T. Ma, L. Huang, Q. Lu, S. Hu, KR-GCN: Knowledge-Aware Reasoning with Graph Convolution Network for Explainable Recommendation, *ACM Transactions on Information Systems* 41 (2023) 4:1–4:27. URL: <https://dl.acm.org/doi/10.1145/3511019>. doi:10.1145/3511019.
- [38] Z. Lyu, Y. Wu, J. Lai, M. Yang, C. Li, W. Zhou, Knowledge Enhanced Graph Neural Networks for Explainable Recommendation, *IEEE Transactions on Knowledge and Data Engineering* 35 (2023) 4954–4968. doi:10.1109/TKDE.2022.3142260, conference Name: IEEE Transactions on Knowledge and Data Engineering.
- [39] S. Geng, Z. Fu, J. Tan, Y. Ge, G. de Melo, Y. Zhang, Path Language Modeling over Knowledge Graphs for Explainable Recommendation, in: *Proceedings of the ACM Web Confer-*

- ence 2022, WWW '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 946–955. URL: <https://dl.acm.org/doi/10.1145/3485447.3511937>. doi:10.1145/3485447.3511937.
- [40] C.-Y. Tai, L.-Y. Huang, C.-K. Huang, L.-W. Ku, User-Centric Path Reasoning towards Explainable Recommendation, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 879–889. URL: <https://dl.acm.org/doi/10.1145/3404835.3462847>. doi:10.1145/3404835.3462847.
- [41] Y. Zhao, X. Wang, J. Chen, Y. Wang, W. Tang, X. He, H. Xie, Time-aware Path Reasoning on Knowledge Graph for Recommendation, *ACM Transactions on Information Systems* 41 (2022) 26:1–26:26. URL: <https://dl.acm.org/doi/10.1145/3531267>. doi:10.1145/3531267.
- [42] T. Silveira, M. Zhang, X. Lin, Y. Liu, S. Ma, How good your recommender system is? A survey on evaluations in recommendation, *International Journal of Machine Learning and Cybernetics* 10 (2019) 813–831. URL: <https://doi.org/10.1007/s13042-017-0762-9>. doi:10.1007/s13042-017-0762-9.
- [43] S. K. Raghuvanshi, R. K. Pateriya, Recommendation Systems: Techniques, Challenges, Application, and Evaluation, in: J. C. Bansal, K. N. Das, A. Nagar, K. Deep, A. K. Ojha (Eds.), *Soft Computing for Problem Solving, Advances in Intelligent Systems and Computing*, Springer, Singapore, 2019, pp. 151–164. doi:10.1007/978-981-13-1595-4\_12.
- [44] B. Abdollahi, O. Nasraoui, Explainable Matrix Factorization for Collaborative Filtering, in: Proceedings of the 25th International Conference Companion on World Wide Web, WWW '16 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2016, pp. 5–6. URL: <https://doi.org/10.1145/2872518.2889405>. doi:10.1145/2872518.2889405.
- [45] G. Peake, J. Wang, Explanation Mining: Post Hoc Interpretability of Latent Factor Models for Recommendation Systems, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 2060–2069. URL: <https://doi.org/10.1145/3219819.3220072>. doi:10.1145/3219819.3220072.
- [46] L. Coba, P. Symeonidis, M. Zanker, Personalised novel and explainable matrix factorisation, *Data & Knowledge Engineering* 122 (2019) 142–158. URL: <https://www.sciencedirect.com/science/article/pii/S0169023X1830332X>. doi:10.1016/j.datak.2019.06.003.
- [47] J. Tan, S. Xu, Y. Ge, Y. Li, X. Chen, Y. Zhang, Counterfactual Explainable Recommendation, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, CIKM '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 1784–1793. URL: <https://doi.org/10.1145/3459637.3482420>. doi:10.1145/3459637.3482420.
- [48] G. Balloccu, L. Boratto, G. Fenu, M. Marras, Post Processing Recommender Systems with Knowledge Graphs for Recency, Popularity, and Diversity of Explanations, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 646–656. URL: <https://dl.acm.org/doi/10.1145/3477495.3532041>. doi:10.1145/3477495.3532041.
- [49] G. Balloccu, L. Boratto, G. Fenu, M. Marras, Reinforcement recommendation reasoning

through knowledge graphs for explanation path quality, *Knowledge-Based Systems* 260 (2023) 110098. URL: <https://www.sciencedirect.com/science/article/pii/S0950705122011947>. doi:10.1016/j.knosys.2022.110098.

- [50] X. Chen, Y. Zhang, J.-R. Wen, Measuring "Why" in Recommender Systems: a Comprehensive Survey on the Evaluation of Explainable Recommendation, 2022. URL: <http://arxiv.org/abs/2202.06466>. doi:10.48550/arXiv.2202.06466, number: arXiv:2202.06466 arXiv:2202.06466 [cs].
- [51] S. Mohseni, N. Zarei, E. D. Ragan, A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems, *ACM Transactions on Interactive Intelligent Systems* 11 (2021) 24:1–24:45. URL: <https://doi.org/10.1145/3387166>. doi:10.1145/3387166.
- [52] S. Chang, F. M. Harper, L. G. Terveen, Crowd-based personalized natural language explanations for recommendations, in: *Proceedings of the 10th ACM Conference on Recommender Systems, RecSys '16*, Association for Computing Machinery, New York, NY, USA, 2016, pp. 175–182. URL: <https://doi.org/10.1145/2959100.2959153>. doi:10.1145/2959100.2959153, 77 citations (Semantic Scholar/DOI) [2022-10-07] 44 citations (Crossref) [2022-10-07].
- [53] F. Costa, S. Ouyang, P. Dolog, A. Lawlor, Automatic Generation of Natural Language Explanations, in: *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion, IUI '18 Companion*, Association for Computing Machinery, New York, NY, USA, 2018, pp. 1–2. URL: <https://doi.org/10.1145/3180308.3180366>. doi:10.1145/3180308.3180366.
- [54] Y. Zhang, X. Chen, Explainable Recommendation: A Survey and New Perspectives, *Foundations and Trends® in Information Retrieval* 14 (2020) 1–101. URL: <http://arxiv.org/abs/1804.11192>. doi:10.1561/15000000066, 297 citations (Semantic Scholar/arXiv) [2022-02-14] 297 citations (Semantic Scholar/DOI) [2022-02-14] arXiv: 1804.11192.
- [55] M. Caro-Martínez, G. Jiménez-Díaz, J. A. Recio-García, Conceptual Modeling of Explainable Recommender Systems: An Ontological Formalization to Guide Their Design and Development, *Journal of Artificial Intelligence Research* 71 (2021) 557–589. URL: <https://jair.org/index.php/jair/article/view/12789>. doi:10.1613/jair.1.12789.
- [56] Z. Fu, Y. Xian, R. Gao, J. Zhao, Q. Huang, Y. Ge, S. Xu, S. Geng, C. Shah, Y. Zhang, G. de Melo, Fairness-Aware Explainable Recommendation over Knowledge Graphs, in: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '20*, Association for Computing Machinery, New York, NY, USA, 2020, pp. 69–78. URL: <https://doi.org/10.1145/3397271.3401051>. doi:10.1145/3397271.3401051.
- [57] G. Vilone, L. Longo, Notions of explainability and evaluation approaches for explainable artificial intelligence, *Information Fusion* 76 (2021) 89–106. URL: <https://doi.org/10.1016/j.inffus.2021.05.009>. doi:10.1016/j.inffus.2021.05.009.